Mercedes Fernández Liporace
Alicia Noelia Cayssials
Marcelo Antonio Pérez

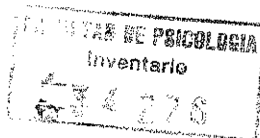
# Curso básico de Psicometría

## Teoría clásica





Fernández Liporace, M. Mercedes; Cayssials, Alicia Noelia; Pérez, Marcelo Antonio
Curso básico de Psicometría : Teoría clásica. - 1a ed. - Buenos Aires: Lugar Editorial, 2009.
176 p.; 20x28 cm
ISBN 978-950-892-335-6
1. Psicología. 2. Psicometría. I. Título
CDD150






ISBN: 978-950-892-335-6

Castro Barros 1754 (C1237ABN) Buenos Aires, Argentina
Tel/Fax: (54-11) 4921-5174 / (54-11) 4924-1555
E-mail: lugared@elsitio.net / info@lugareditorial.com.ar
www.lugareditorial.com.ar

Queda hecho el depósito que marca la ley 11.723
Impreso en la Argentina – Printed in Argentina




# *Introducción*

El objetivo de este libro es el de acercar al alumno de la Carrera de Psicología de nuestro medio a algunos de los temas básicos de la Psicometría y de la Evaluación Psicológica, de una manera más acorde con la instrucción que han recibido durante las asignaturas ya cursadas, y teniendo como meta de largo plazo la formación de un profesional capacitado para desempeñarse en el ámbito de aplicación de la evaluación, en tanto usuario responsable de las herramientas con que se cuenta en esta área de la Psicología. Este texto, así, puede convertirse en el inicio de una lista de obras de consulta destinadas a lograr una introducción a los conceptos de estas disciplinas y, a la vez, servir de guía preliminar para elaborar un plan de especialización ulterior.

A pesar de la existencia de diversos manuales traducidos al castellano o escritos originalmente en lengua española, ninguno de ellos parece adaptarse completamente a las necesidades de los estudiantes que se forman en nuestro país, así como tampoco a los planes de estudio vigentes en la mayoría de nuestras facultades de Psicología. Es por ello que un grupo de profesores con años de experiencia en investigación psicométrica, en el dictado de esta materia, de asignaturas afines y de cursos de posgrado relacionados, hemos decidido escribir una obra accesible pero suficiente para lo que esperamos de nuestros alumnos al aprobar el curso de psicometría y evaluación psicológica: un colega con criterio formado, capaz de elegir las técnicas más apropiadas según el objetivo de una evaluación dada, conforme a los estándares de calidad vigentes internacionalmente en nuestra disciplina en la actualidad y acorde a las características de la/s persona/s o grupo/s evaluado/s y de la situación o ámbito de aplicación en que tal proceso se dé; un colega que pueda encontrarse con una técnica de evaluación que no ha conocido durante la cursada y, sobre la base de lo aprendido, sea potencialmente capaz de lograr utilizarla por sus propios medios, habiendo accedido a la posibilidad de *aprender a aprender*; un colega que sea consciente de cuánto sabe sobre el tema y de cuánto ignora, para que continúe formándose y estudiando por su cuenta, y también un colega que tenga la suficiente responsabilidad para derivar aquellos casos que no está capacitado para atender. O bien un psicólogo no especializado en Evaluación, pero que pueda convertirse en avezado lector de informes psicológicos elaborados por especialistas, y sea capaz de valorar adecuadamente la pertinencia y validez de las afirmaciones volcadas en esos documentos. Si tales objetivos se logran, nuestro propósito estará cumplido.

Y entendemos que a estas metas se accede mediante una formación teórica y práctica organizada con responsabilidad, construida a partir de la absorción de conocimientos, sustentada en una permanente actitud de crítica y cuestionamiento y, por supuesto, fundamentándose en una postura ética de base.

Otro de los propósitos que este texto acomete se relaciona con dejar en claro que la Evaluación Psicológica y las técnicas psicométricas tienen limitaciones claras y jamás aportan al especialista los mejores instrumentos de que éste dispone: su *expertise* y su sentido común. Estas herramientas psicométricas pierden sentido si no se incluyen en un contexto más amplio de evaluación, con determinados propósitos, y que se desarrolla en unas coordenadas dadas de tiempo y espacio, siendo aplicadas a personas o a grupos con características específicas. Si perdemos de vista ese norte, la tarea profesional quedará desvirtuada, con todos los riegos que ello implica.

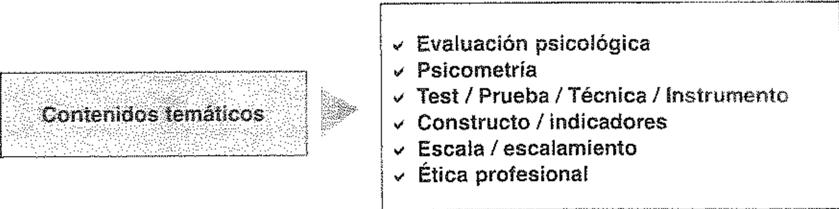
Esperamos que este libro sea un pequeño aporte en una dirección positiva.

Asimismo, queremos agradecer y recordar a la Dra. María Martina Casullo –recientemente fallecida–, que ha sido una de las precursoras al introducir en nuestro país una visión más actual de esta área de trabajo, construyendo un equipo del que formamos parte, que se desempeña en las áreas de docencia e investigación y que sigue creciendo con la incorporación y entrenamiento constante de nuevos recursos humanos.

*Capítulo 1*

# *Psicometría, evaluación psicológica y ámbitos de aplicación*

Mercedes Fernández Liporace




## 1.1 La Evaluación Psicológica: concepto y caracterización

Frecuentemente, cuando las personas se encuentran ante una situación de evaluación psicológica experimentan diversos sentimientos y emociones, tales como ansiedad, temor, indiferencia, curiosidad, hastío, aburrimiento, entre otras posibles. Esto sucede porque cuando un individuo sabe que está siendo evaluado desde el punto de vista psicológico, suele entregar al evaluador un material muy precioso: su desempeño ante un test de inteligencia o de habilidades, sus respuestas sobre temas personales que se le han planteado, sus preferencias, sus intereses, sus actitudes, sus comportamientos más frecuentes, sus síntomas, entre otros atributos pasibles de ser examinados.

Es así como ello suele despertar en el evaluado cierta ansiedad acerca de los resultados; el *saber cómo le fue* tiene que ver, fundamentalmente, con dos cuestiones: la primera se vincula con que ha brindado determinados datos de su historia, ha puesto en juego ciertas aptitudes, ha revelado partes de su vida cotidiana o de sus comportamientos más habituales y todo ello se relaciona con *temas muy personales*; la segunda se refiere a que esta persona sabe que está siendo evaluada en cuanto a estos contenidos en una situación dada, con un objetivo en especial y con el propósito de tomar una decisión determinada en un cierto sentido: en base a esos resultados se recomendará que quien ha sido sujeto de la evaluación sea o no incluido en un programa de entrenamiento especial, sea derivado o no a una psicoterapia, sea seleccionado o no para un puesto de trabajo –o promovido o no a una posición de mayor jerarquía–, sea declarado imputable o inimputable de algún delito que presuntamente haya cometido, sea valorado como competente o no para hacerse cargo de la crianza o tutela de algún niño, entre otras muchas circunstancias que pueden darse. El

sujeto *sabe* –intuitivamente o porque se lo han explicado– que a partir de esta evaluación psicológica se recomendará un curso de acción específico, se tomará alguna decisión que, directa o indirectamente, afectará su vida y/o la de sus allegados en mayor o menor grado.

A partir de lo previamente explicitado, puede claramente percibirse que ***aquello que llamamos Evaluación Psicológica no es ni más ni menos que un proceso de toma de decisiones*** (Cronbach & Gleser, 1957), puesto que el propósito último de la misma es recomendar un camino de acción determinado en virtud de los objetivos perseguidos por la evaluación. Debe quedar claro, entonces, que no es el psicólogo evaluador quien ha de tomar esta decisión, sino que sólo efectuará una recomendación; es así que puede caracterizarse este proceso como una instancia consultiva, no decisoria, pero que a la larga implicará la puesta en práctica de una decisión tomada por otras personas distintas del evaluador.

Si tenemos en cuenta que la Psicología como disciplina aplicada se desarrolla principalmente en los ***contextos clínico, laboral, educativo y forense***, entre otros, la Evaluación Psicológica como subdisciplina englobada en la Psicología aplicada, también tiene lugar en los mismos ámbitos antes nombrados. Por ejemplo, en el ***ámbito forense***, será el juez quien determine el curso de acción a seguir en virtud de las recomendaciones que el psicólogo ha redactado; el juez ha solicitado determinados puntos de pericia, por caso, determinar si un sujeto era capaz de comprender o no la criminalidad de un acto al momento de cometerlo, o si alguien que ha sido víctima de un accidente automovilístico padece un trastorno por estrés post-traumático. Luego, el psicólogo evaluador aportará indicadores acordes con los puntos de pericia y elaborará un informe en el que detallará todos estos elementos, que fundamentarán sus conclusiones y recomendaciones, que serán elevadas al juez. Con todos estos elementos, el magistrado tomará una decisión concreta, teniendo en cuenta lo reportado por el perito psicólogo que se ha encargado del proceso de evaluación –llamado pericia en el ámbito forense–.

En el ***ámbito educativo***, en cambio, quien solicita esta evaluación podría ser un maestro, un director de escuela, un papá, una mamá, entre otros actores posibles. Los motivos que llevan a efectuar este pedido podrían ser, por ejemplo, disfunciones en el aprendizaje, trastornos en el comportamiento dentro de la escuela, dificultades de integración en un grupo de pares o en la comunicación con algún docente, detección de talentos, o evaluación de habilidades especiales, entre otros. Con el motivo de consulta como guía, el psicólogo evaluador reunirá toda la evidencia posible que lo lleve a poder formular un diagnóstico de la situación y una o varias recomendaciones alternativas, dirigidas a solucionar o mejorar la dificultad que motiva la evaluación, o a describir y recomendar intervenciones o programas acordes con los talentos o habilidades especiales detectados. Luego, será la escuela –como institución–, el docente, el psicopedagogo del gabinete escolar, la familia del niño o todos los actores involucrados en la problemática, los que deberán adoptar o no esa decisión de la que hasta aquí se habla, pero que simplemente ha sido *sugerida* por el profesional a cargo del proceso de evaluación.

Una situación semejante se da en el ***ámbito laboral***: el psicólogo es convocado para efectuar una recomendación, en relación con la promoción o selección de un candidato entre varios potenciales, o con la detección de áreas de vacancia que deben fortalecerse en un grupo de trabajo determinado para diseñar un plan de capacitación específico. Pero es la empresa contratante –representada en la o las figuras de sus gerentes o personal en roles ejecutivos de jerarquía– quien tomará la decisión última de optar o no por obrar según lo recomendado por el evaluador. Este curso de acción ha sido sugerido como resultado de la acumulación de una serie de indicadores que, tomados en su conjunto, dan respuesta a determinadas preguntas que se relacionan con el objetivo mismo de la evaluación.



En el ***ámbito clínico***, donde para algunos autores el proceso de evaluación psicológica se asimila al nombre de *psicodiagnóstico* (Casullo, 1996), el motivo de consulta está gatillado por algún malestar, sufrimiento, disfunción, sintomatología o por algún grado de invalidez constatado en un individuo, pareja, familia o grupo. Esta disfunción puede adoptar diferentes formas e implicar distintos grados de gravedad o compromiso y puede haber sido puesta en evidencia por el propio sujeto que la padece o por alguno de sus allegados. Con este motivo de consulta inicial, el evaluador recogerá toda la información posible vinculada al tema, de manera de poder identificar alguna o algunas recomendaciones, dirigidas a mejorar la situación presente. Pero será el propio evaluado y/o sus allegados quienes decidirán, concretamente, si seguirán las indicaciones dadas por el psicólogo.

Más allá de la visión clínica tradicional, actualmente está tomando mucha fuerza el enfoque proveniente del paradigma de la Psicología Positiva, que propende a destacar y enfatizar los factores protectores con que las personas cuentan: en tanto algunos individuos enferman o agravan sus patologías ante determinadas circunstancias, otros salen incólumes e incluso, fortalecidos de las mismas situaciones. Así, sin perder de vista los componentes disfuncionales, la detección de aspectos salugénicos –generadores o coadyuvantes de la salud– implicados en los niveles de análisis individual, grupal y social se vuelve cada día una tendencia más pronunciada en el estilo de trabajo de los profesionales dedicados a la evaluación, proveniente de este viraje teórico en la manera de abordar el objeto de estudio (Aspinwell & Staudinger, 2003; Casullo, 2003; Henderson & Milstein, 2003; Linley & Joseph, 2004; Maddux, 2002; Peterson & Seligman, 2004; Seligman, Steen, Park & Peterson, 2005; Snyder & Lopez, 2002)

Por último, existe otro ámbito de aplicación en el que pueden efectuarse tareas de evaluación psicológica, y es la ***evaluación de programas***; en esta área de trabajo la actividad del evaluador está encaminada a determinar la eficacia de una intervención, tratamiento o programa; por ejemplo, la eficacia que ha tenido una campaña dirigida a modificar las actitudes hacia la integración de personas con capacidades especiales, o a cambiar los comportamientos referidos al uso de preservativos en los individuos sexualmente activos, o a mejorar la imagen de un candidato perteneciente a un partido político, o una campaña orientada a introducir un nuevo producto en el mercado adolescente, o una intervención diseñada para elevar el desempeño en lectura expresiva de niños de segundo grado, o a modificar determinadas interacciones grupales con *bullying* –en las que se produce un interjuego de relaciones en las que un grupo de niños o adolescentes victimiza a otro grupo o individuo física o psíquicamente–, o un tratamiento para dejar de fumar con la menor ansiedad posible. En todos estos casos se habla de una intervención, programa o tratamiento orientado a cambiar ciertas circunstancias que se desea alterar. Para ello debe efectuarse, *al menos*, una evaluación de la situación *antes* de aplicar el programa y otra *después* de haberlo concluido, con miras a determinar si éste resultó eficaz, en términos de lograr la modificación pretendida y en el sentido esperado.

Tomando el último de los ejemplos antes enumerados, debería determinarse, por caso, la cantidad de cigarrillos que los participantes en el tratamiento fumaban habitualmente antes de iniciar el mismo, acompañándose, por ejemplo, de la medición de su nivel de ansiedad; una vez aplicado el programa sería preciso relevar el número de

cigarrillos que terminaron consumiendo después de aquél y el monto de ansiedad correlativo; para que tal intervención sea calificada como eficaz, la cantidad de cigarrillos consumidos post-tratamiento debería ser significativamente menor que los que fumaba en el pre-tratamiento la mayoría de los participantes, a la vez que la ansiedad debería resultar, al menos, igual a la registrada pre-tratamiento, pero no mayor. También podrían introducirse otras variantes tales como hacer un seguimiento a los seis meses y al año de finalizar el programa –aunque además podría incluirse alguna medición al promediar el mismo– e incorporar un grupo de comparación o control que no haya participado del tratamiento, para estar seguros de que la disminución en la cantidad de cigarrillos ha sido significativamente mayor en el grupo que asistió al programa que en el que no participó de él, al tiempo que no se han producido aumentos significativos en la ansiedad del grupo incluido en la intervención respecto del grupo control. Este ejemplo se introduce aquí ya que resulta de fácil comprensión, aunque cuando se habla de evaluación psicológica de programas, el lector debe imaginarse la posibilidad de trabajar también con escalas para evaluar actitudes –si es que se trató de introducir modificaciones en las mismas–, con tests de habilidades varias –si se dirigió la intervención a mejorar una destreza, aptitud o competencia dada–, con sociogramas –si lo que se intentó fue facilitar las interacciones entre los miembros de un grupo– o con inventarios o check-lists (listados) de comportamientos –si se pretendía, por ejemplo, incorporar nuevos repertorios conductuales–.

La determinación de la eficacia de un programa, en definitiva, es un problema de economía privada o pública, ya que si se verifica que aquél no ha resultado todo lo exitoso que se esperaba, lo conveniente sería decidir su modificación, suspensión o no repetición, reemplazándolo por otra alternativa que luego se evaluará en cuanto a su eficacia específica.

Habiendo puesto el foco en los ámbitos de aplicación en los que el psicólogo evaluador puede desempeñar su tarea, también debe advertirse que, en muchas ocasiones, se hace evaluación cuando se recogen datos para investigación. En este sentido, teniendo presente que la Evaluación Psicológica tiene cabida en diversos ámbitos de aplicación de la Psicología –clínico, educativo, laboral, forense, evaluación de programas–, también puede decirse que esta disciplina se ejerce en el ámbito de investigación, más específicamente, en los contextos de descubrimiento y de justificación –es decir, tanto cuando los fenómenos se describen cuanto en el momento en que se descubren, identifican o formulan leyes, así como cuando se justifican mediante evidencia empírica aportada por datos de la realidad– (Fig. 1.1).

Resumiendo, en el ámbito de aplicación, en tanto proceso de toma de decisiones, entonces, la Evaluación Psicológica implica una instancia consultiva, en la que el psicólogo será convocado para reunir toda la información o indicadores disponibles referidos a un problema, a una pregunta o tema dado, y en base a ello efectuará una o varias recomendaciones dirigidas a resolver o mejorar el problema o a responder el interrogante aún sin resolución, en la que la decisión estará en manos de otra persona –quien ha solicitado la evaluación–. En cambio, si se trabaja en el ámbito de investigación, la tarea de evaluación se relacionará con recabar datos o indicadores que den cuenta de aquel fenómeno que se quiere investigar, de manera que esta actividad de recogida de datos será un paso más dentro del proceso de investigación.

Entendemos por ámbitos de aplicación de la Psicología y de la Evaluación Psicológica, aquellos contextos en los que puede desempeñarse el psicólogo –en este caso, el psicólogo evaluador–, ejerciendo su rol profesional, aplicando conocimientos técnicos que derivan de teorías y de investigaciones empíricas, es decir, del ámbito de la



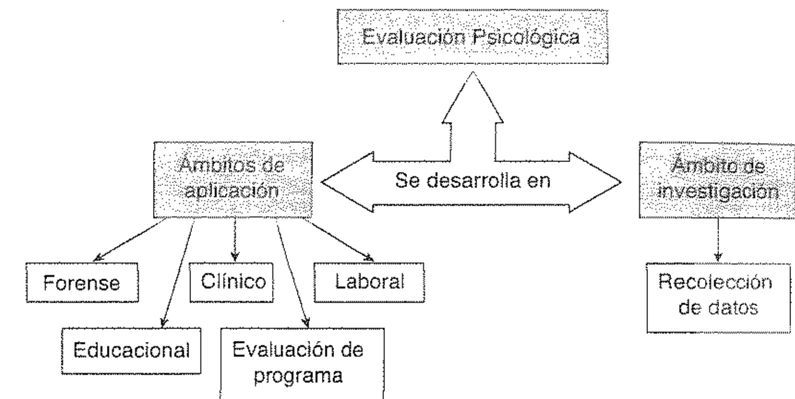



Fig. 1.1 Evaluación Psicológica en los ámbitos de aplicación e investigación

investigación científica a la resolución de situaciones o problemas concretos; estos ámbitos son, entonces, el clínico, el educacional, el forense, el laboral y el de evaluación de programas.

Por otra parte, definimos como ámbito de investigación al contexto en el que la Psicología se desarrolla como disciplina científica, generando nuevos conocimientos en forma permanente; esos conocimientos serán transferidos, mediata o inmediatamente, al ámbito de aplicación, en beneficio de quienes se sirvan de aquéllos para solucionar ciertas dificultades o para mejorar situaciones o estados concretos, en su papel de profesionales especializados en el área. De todo lo anterior puede apreciarse, fácilmente, el permanente interjuego entre los ámbitos de aplicación y de investigación de la Psicología, y más específicamente, de la Evaluación Psicológica.

Así, podemos afirmar que, en líneas generales y tomando en cuenta el trabajo profesional del psicólogo, la Evaluación Psicológica es una tarea de la psicología aplicada dirigida a la solución de problemas personales, institucionales, grupales, comunitarios, sociales o ambientales. Avanzando más en este concepto, podemos agregar que para desarrollar esta actividad resulta necesaria la aplicación de un modelo teórico que será de utilidad para comprender o analizar el fenómeno concreto que es objeto de nuestra atención. Ello significa que entre las tareas implicadas en la Evaluación Psicológica se encuentran la categorización, la comparación, el análisis y la contrastación de datos referidos a atributos del sujeto y/o de la situación o interacción que se está analizando (Anguera, 1995; Casullo, 1996; Fernández Ballesteros, 1993; Forns i Santacana, 1993; Silva, 1990). Por todo lo antedicho, resulta sencillo colegir que el psicólogo evaluador podrá posicionarse en diversos modelos teóricos alternativos para *"leer"* e interpretar el fenómeno que está estudiando o analizando, no existiendo un único marco posible que sustente las conclusiones a las que se arribe. De esta manera, efectuando una clasificación sencilla, se pueden distinguir tres tipos principales de modelos en los que podemos basarnos: los centrados en el sujeto, los que ubican el énfasis en las variables situacionales y los que colocan el acento en las interacciones entre sujeto y medio (Casullo, 1996; Forns, Kirchner & Torres, 1991).

Dentro de los modelos centrados en el sujeto, podemos localizar, por ejemplo, el médico-psiquiátrico, el psicoanalítico, el de los rasgos o atributos, y el fenomenológico, entre los más importantes.

En el grupo de los modelos teóricos que enfatizan las variables situacionales encontramos el conductismo radical, los mediacionales y, fundamentalmente, aquellos que jerarquizan la influencia de variables intervinientes.

Por último, en la categoría correspondiente a los marcos teóricos que privilegian el análisis de las interacciones entre el sujeto y el contexto, aparecen los modelos interaccionales, de condicionamiento, el estructuralismo cognitivista, el de las representaciones sociales, el estudio del afrontamiento y del procesamiento cognitivo, entre los más representativos (Fig. 1.2.).

Desde este punto de vista, ampliando la definición básica de la Evaluación Psicológica que antes examinábamos, que la caracteriza como un *proceso de toma de decisiones*, podría agregarse, avanzando un paso más, que "[...] *el objeto de esta área de trabajo es el estudio, análisis y valoración de las características de un sujeto, de sus formas de acción, reacción e interacción con los demás y con la realidad, y de sus procesos de cambio*" (Forns i Santacana, 1993, p. 20). Para llegar a esta meta es preciso conceptualizar al individuo o al grupo como integrante de un sistema conformado por sujetos caracterizados como individualidades bio-psico-sociales, sometidas a procesos internos y externos que afectan y determinan recíprocamente los modos de contacto entre el sujeto y la realidad; los efectos de estos contactos se manifiestan en producciones o comportamientos de tipo fisiológico, motriz, emocional y/o cognitivo, en tanto que la totalidad de los factores que influyen se hallan constreñidos por mecanismos de interrelación recíproca. Según esa autora, actualmente resulta imposible pensar la Evaluación Psicológica como una actividad que se reduce meramente al análisis de *productos o comportamientos* separados de otras variables; asimismo también se vuelve dificultoso valorar los *procesos* en forma aislada, sin relación con el sujeto concreto que los ha elaborado y en función de determinadas variables situacionales. De esta manera cada día se trabaja más desde una perspectiva integradora y holística que intenta no dejar de lado ninguno de todos los aspectos mencionados. Maganto, Amador y González (2001) caracterizan la forma de actuar del psicólogo en la tarea de evaluación como integrada por varios pasos que no necesariamente se disponen de manera secuencial sino que pueden darse, por momentos, en un desarrollo simultáneo y/o sucesivo; así, el especialista en el área *recabará información mediante diversas herramientas técnicas* que seleccionará cuidadosamente según el objetivo de la evaluación, *organizará dicha información en un mapa conceptual* que contemple la temporalidad y pluricausalidad de los fenómenos psíquicos, *formulará hipótesis diagnósticas y explicativas* acerca de la naturaleza y definición de los procesos y productos involucrados en la evaluación, que irá fortaleciendo o descartando según avance en el análisis del material recogido y contraste dichas hipótesis con datos de la realidad; elaborará una *síntesis del caso* y efectuará una o varias *recomendaciones* en términos de cursos de acción sugeridos, que comunicará mediante una *devolución oral y/o* un *informe*. Finalmente, de ser posible, realizará una *evaluación de control, seguimiento o análisis de cambio*, según corresponda, efectuados con el objeto de validar en la práctica los resultados de su evaluación.



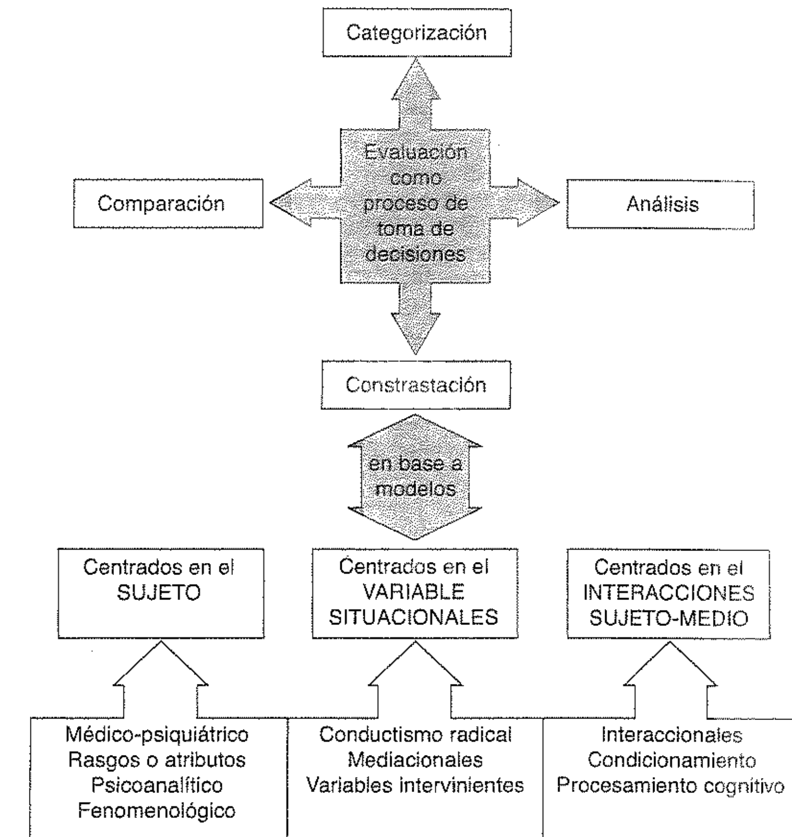



Fig. 1.2. Modelos teóricos en los que puede basarse la Evaluación Psicológica

Dejando ahora las teorías, y desde la arista técnica, para llevar a cabo su tarea de evaluación, el psicólogo especialista en el área dispone de un arsenal de pruebas que le permiten recoger la información necesaria para llevar a cabo este proceso. Ellas son, haciendo una clasificación preliminar y superficial, las *técnicas psicométricas, las proyectivas* y, por supuesto, las *entrevistas*, entendidas estas últimas como elemento técnico irreemplazable en la toma de contacto directo con el entrevistado y sus aspectos particulares, así como con el propósito de la evaluación (Albajari, 1996; Avila Espada, 1989; Nahoum, 1961; Siquier de Ocampo, García Arzeno & Grassano, 1987; Rolla, 1981; Rogers, 1966; Sullivan, 1959). La entrevista constituye una categoría aparte dentro de los tipos de instrumentos de evaluación existentes; sin embargo, en sentido amplio, dependiendo de la clase de entrevista de que se trate, podría intentarse una clasificación de la misma en el grupo de las herramientas psicométricas o proyectivas, como veremos más adelante.

## 1.2 Evaluación Psicológica y Psicometría: diferencias e interacción

En el apartado anterior nos referíamos a la Evaluación Psicológica como una subdisciplina ubicada dentro de la Psicología que puede desarrollarse dentro de los ámbitos de aplicación o de investigación; en el primero de los casos, el psicólogo asume el rol de usuario de técnicas, aplicador, administrador o evaluador –debe recordarse que estos rótulos simplemente se utilizan para marcar las diferencias en el papel del profesional que trabaja en el área en diferentes ámbitos, pero la idea es que el usuario de técnicas no se transforme en un mero *testista* o *aplicador* de instrumentos, sino que sea un especialista formado, con una amplia experiencia práctica y una entrenada capacidad interpretativa sustentada por un extenso bagaje teórico–. En el segundo caso, cuando el psicólogo se desempeña en el ámbito de investigación –tanto en el contexto de descubrimiento como en el de justificación–, el principal objetivo de su tarea es la generación de nuevos conocimientos que, en última instancia, serán transferibles al ámbito de aplicación de la Psicología para la resolución de problemas concretos. Es entonces que en esta circunstancia el especialista se desempeñará como investigador, utilizando las técnicas de evaluación psicológica como instrumentos e recolección de datos, orientados a obtener la información empírica –proveniente de la realidad– que servirá para corroborar o refutar las hipótesis de trabajo que se hubieren formulado a la luz de un modelo teórico dado.

Ahora bien, existe otra inserción posible para el psicólogo investigador, y ella es la de especializarse en el diseño, construcción y adaptación de técnicas de evaluación psicológica; en este caso, estas herramientas no son el medio utilizado para lograr un fin, como sucede en el ámbito de aplicación, en el que las técnicas se usan para recabar información sobre un sujeto o grupo respondiendo a un motivo de consulta o a un objetivo de evaluación dado, o como acontece cuando en el ámbito de investigación las técnicas son medios para recoger datos para llevar a cabo un estudio empírico. En la circunstancia señalada en un principio, el investigador se especializará en el área de la *psicometría*, donde los tests ya no serán un medio para obtener información sino un fin en sí mismos; ellos son, en este caso, el producto de un desarrollo tecnológico derivado de un modelo teórico y su construcción será, en sí misma, un objetivo de investigación que más tarde servirá para que los profesionales que trabajan en el ámbito de aplicación de la evaluación utilicen en su quehacer diario (Fig. 1.3.).

Siguiendo esta línea de razonamiento, se define a la ***Psicometría*** como *la disciplina que tiene por finalidad el desarrollo de modelos –preferentemente, pero no de manera exclusiva– cuantitativos que permitan "transformar" o codificar los fenómenos o los hechos en datos, diseñando métodos adecuados para la aplicación de tales modelos con el fin de determinar las diferencias individuales de los sujetos en cuanto a sus atributos, sus propiedades o sus rasgos* (Cliff, 1973; Martínez Arias, 1995). A pesar de que tradicionalmente se ha entendido la Psicometría como la disciplina que se ocupa de la construcción de pruebas, esta atribución es una consecuencia del primer objetivo nombrado, que se refiere a la generación o formulación de modelos psicométricos. Las pruebas son, simplemente, un desarrollo tecnológico derivado de los modelos teóricos de corte psicométrico. Sin embargo, popularmente, se conoce a la Psicometría más como una disciplina utilitaria y tecnológica, antes que como una rama teórica de la Psicología.

Es importante destacar que los modelos psicométricos formulados para explicar o comprender fenómenos –siempre referidos a las diferencias individuales entre las



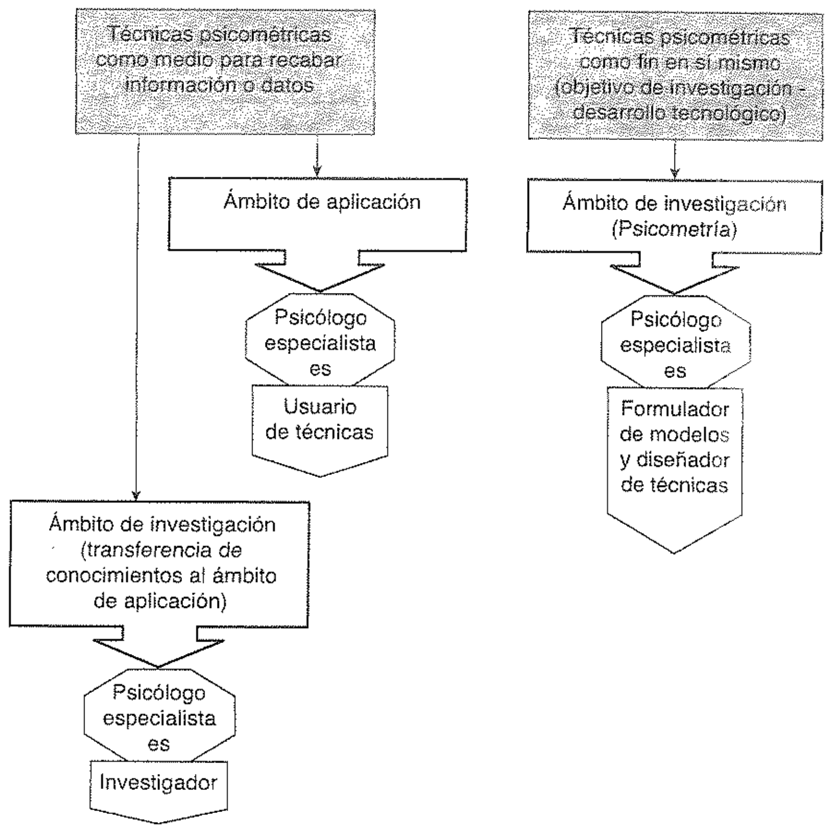



Fig. 1.3. Rol del psicólogo especialista en relación con las técnicas psicométricas según su ámbito de trabajo

personas– son, en general, de corte cuantitativo, pero no necesariamente esto es así en todos los casos.

Es posible detectar y describir diferencias entre los sujetos mediante el empleo de conceptos operacionalizados en forma cualitativa, semicuantitativa u ordinal, o directamente cuantitativa, tal como expresábamos en el párrafo precedente. Para comprender estas distinciones resulta útil apelar a la tradicional clasificación de los niveles de medición que debemos a Stevens (1946), que divide a las escalas de medición en nominales, ordinales, de intervalos y de razones. Sin embargo, previamente a desarrollar estas categorías, es preciso caracterizar la *medición* como un proceso de asignación de números o símbolos a atributos de los objetos –sujetos, en el caso de la Psicología– siguiendo una serie de reglas dirigidas a representar la presencia o ausencia del atributo medido, su cualidad, jerarquía o cantidad. Conceptualizada de esta forma, una escala se define como un conjunto de números o símbolos cuyas propiedades *modelan* propiedades empíricas de los sujetos a los que esos números o símbolos

son asignados (Cohen & Swerdlik, 2001). Es importante tener en cuenta que esta caracterización permite diferenciar distintas maneras de medir, abandonando la postura más conservadora, que propone la sinonimia entre medición y cuantificación. Siguiendo esta línea de pensamiento, medimos cuando decimos que un trozo de camino tiene una longitud de 7 km, cuando expresamos nuestra edad en años, cuando contamos la cantidad de monedas que tenemos en nuestro bolsillo, la cantidad de veces que un niño deambulador se dirige a su madre, cuando tomamos nota de la cantidad de respuestas correctas que un adolescente brinda en un test de vocabulario, del tiempo que tarda un examinado en elaborar una respuesta, o la cantidad de errores que un escolar comete en una prueba de madurez visomotriz; en todos estos casos utilizamos la manera más corriente que tenemos de pensar la medición, como sinónimo de cuantificación, igual a cantidad de unidades que tomemos como referencia para realizar la medición –km, años, cantidad de monedas, número de acercamientos, cantidad de respuestas correctas, tiempo de reacción, número de errores cometidos–.

Ahora bien, en este nivel cuantitativo o métrico pueden, a su vez, distinguirse las escalas de intervalos y las escalas de cocientes o razones. Si bien ambas implican la noción de cuantificación (establecer la cantidad del atributo que el sujeto posee), conservan algunas diferencias. El nivel intervalar, como su nombre lo indica, supone la igualdad de intervalos entre números: cada unidad de esta escala es exactamente igual a cualquier otra unidad; además, al igual que en el nivel de medición ordinal, no existe un punto cero absoluto, sino uno arbitrario. La temperatura expresada en grados centígrados suele ser uno de los ejemplos clásicamente citados en este punto. El punto 0ºC no indica ausencia del atributo temperatura sino que es un hito convencional que resulta cómodo de identificar, ya que es el valor de temperatura en que el agua cambia de un estado a otro, pero está claro que por debajo de los 0ºC sigue habiendo temperaturas, ya que de hecho, existen valores inferiores a esa marca. Muchas escalas para medir la inteligencia utilizan el CI (coeficiente intelectual) como puntaje transformado en el que se expresa el rendimiento de los examinados, que es una escala de intervalos; de acuerdo con lo antedicho, la diferencia que existe entre el intervalo que va de los CI 85 a 100 es similar a la que existe entre los CI de 100 y 115. Por esta razón es posible efectuar en este nivel de medición todas las operaciones aritméticas posibles y, por lo tanto, calcular todos los descriptivos previstos por la Estadística, cosa que no sucede en los niveles de medición ordinal y nominal; por ejemplo, no puede calcularse la media o promedio de atributos tales como acierto/error, síntoma presente o ausente, tipo de neurosis (nivel nominal o cualitativo), gravedad del trastorno psicótico –leve, moderado, severo–, nivel de satisfacción del sujeto con respecto a su trabajo –alto, medio o bajo– (nivel ordinal), ya que esos valores de la variable no pueden sumarse y dividirse para ser promediados. Sí puede calcularse el promedio de calificaciones logradas por los alumnos en un curso (se suman todas las calificaciones y se dividen por el número de alumnos) o la media de las puntuaciones de CI obtenidas por una muestra de sujetos, ya que las unidades de los intervalos de CI, como explicamos antes, son iguales. Por otra parte, en base a razones éticas, el 0 en la escala de CI se ha establecido en niveles de desempeño tan bajos que resulta imposible lograrlo (-5 probables errores respecto de la media), por lo que claramente puede apreciarse la arbitrariedad de ese punto cero establecido.

El nivel de medición de cocientes o razones posee todas las características y facilidades nombradas para al nivel de intervalos, pero su punto cero no es arbitrario sino absoluto; ello significa que no ha sido convencionalmente establecido, sino



que verdaderamente implica ausencia del atributo medido. Así, una persona puede responder que tiene 2, 1, 4 hijos o ningún hijo (cero), queriendo decir este valor que carece realmente de hijos. Lo mismo sucede, por ejemplo, con el dinero –podemos tener $0 en la billetera–, la cantidad de errores que se cometen en un test de maduración visomotriz –es posible tener un desempeño perfecto en una prueba de este tipo–. Como se verá, este nivel de razones implica, al igual que el de intervalos, cuantificación, la posibilidad de efectuar todas las operaciones aritméticas y unidades iguales, pero agrega la distinción de poseer un cero absoluto que indica ausencia real del atributo, en tanto que en el nivel de intervalos, el cero es arbitrariamente fijado.

Más allá de la cuantificación, también medimos cuando asignamos números de tal forma que éstos reflejen un ordenamiento o jerarquía en los sujetos, resultante de la aplicación de algún criterio clasificatorio. Este escalamiento *ordinal, jerárquico o semicuantitativo*, en el que el número es utilizado para identificar cada categoría, se usa para dar cuenta del ordenamiento o posición del sujeto en el rasgo medido, sin implicar una unidad de medición. De esta manera, no se conoce la cantidad absoluta del atributo, sino que solamente se puede establecer qué sujetos se ubican en la misma categoría, qué otras personas caen en la categoría superior o inferior y qué categoría implica más del atributo –en términos generales– respecto de cuál otra. Así, por ejemplo, la variable nivel educativo podría medirse según una escala ordinal cuando asignamos el número 1 a la categoría que indica el más bajo nivel de instrucción, como puede ser "primaria incompleta", que se asigna a todas las personas que, independientemente de cuántos años de ese nivel hayan cursado, no lo han finalizado. La categoría 2 nombrará a quienes tengan una escolaridad primaria completa; la 3, secundaria o media incompleta. 4 y 5 corresponderán a secundaria completa y terciaria/universitaria incompleta, respectivamente, mientras que la 6 se asignará a quienes hayan concluido el nivel terciario/universitario. De esta manera, es fácil apreciar que los sujetos englobados en la categoría 3 (nivel medio incompleto) superan en cantidad de cursos aprobados a los que pertenecen a la 2 (nivel primario completo), pero no puede saberse exactamente qué cantidad de años cursó quien tiene su escuela secundaria inconclusa; de hecho, se ubicarán en esa categoría tanto aquellos que sólo han cursado un año de ese nivel cuanto quienes hayan llegado casi a aprobar el último año pero adeuden una sola materia. Así, en ejemplos como el anterior, resulta sencillo advertir que este nivel de medición permite efectuar un ordenamiento de los sujetos sin que sea posible establecer la cantidad absoluta del atributo medido. Por caso, la manera cuantitativa (nivel de razones) de medir la escolaridad podría ser contando la cantidad de años de educación formal aprobados.

Otro ejemplo del nivel ordinal al que hacemos alusión se pueden observar al clasificar la gravedad de un episodio depresivo mayor en leve, moderado o severo; esta categorización, como decíamos, no implica una cuantificación sino un ordenamiento del atributo *gravedad del trastorno* según un criterio ordinal o semicuantitativo, ya que el límite para dividir las categorías no está dado por ninguna puntuación o cantidad sino por determinadas expresiones sintomáticas que deben ser establecidas con precisión por quien elabora la clasificación; sin embargo, a pesar de ello, dentro de una misma categoría entrarán personas con el mismo estatus de gravedad, pero distintos matices sintomáticos. Más ejemplificaciones de este tipo con variables psicológicas pueden verse en los ordenamientos por edad que pueden hacerse de los adultos según un criterio evolutivo (adultos jóvenes, adultos medios, adultos maduros y adultos mayores); si bien las edades de corte para ubicar a cada individuo en una categoría deben ser establecidas en forma precisa, una vez clasificados los sujetos en cada

una de ellas, no podemos conocer su edad absoluta –cantidad del atributo– sino sólo los rangos de edad entre los que este valor puede variar. Otro caso se da cuando en las escalas de actitudes, intereses, comportamientos o inventarios de personalidad se brinda un formato de respuesta *likert*, que ordena las opciones para responder según este criterio jerárquico no cuantitativo, en el que se solicita al examinado que elija su respuesta dentro de un gradiente de alternativas posibles, que indicarán su grado de conformidad o la frecuencia con la que se presentan los fenómenos o comportamientos descriptos en la formulación del ítem (ver Fig. 1.4.). Tanto el nivel ordinal cuanto el intervalar suponen el establecimiento de un orden y el uso de intervalos; aunque la diferencia radica en que en el ordinal los intervalos no son necesariamente iguales ni suponen una única unidad de medida, en tanto que en el intervalar sí nos hallamos ante intervalos iguales con igual unidad de medida.

**Marque con una cruz o tilde la respuesta que mejor describe la situación en la que Ud. se encuentra más habitualmente:**

Mi pareja me apoya en las actividades que decido emprender

☐ Siempre ☐ A veces ☐ Nunca

**Indique su parecer marcando la opción que considere más adecuada en su caso:**

Generalmente estoy conforme y feliz con la vida que llevo

| ☐ | ☐ | ☐ | ☐ | ☐ |
|---|---|---|---|---|
| Totalmente de acuerdo | En parte de acuerdo | Ni de acuerdo ni en desacuerdo | En parte en desacuerdo | Totalmente en desacuerdo |

**Lea estas afirmaciones y responda indicando su parecer señalando la respuesta que corresponda:**

Creo que la forma en que me enfrento a situaciones de tensión es

| ① | ② | ③ | ④ |
|---|---|---|---|
| Sumamente adecuada | Adecuada | Poco Adecuada | Inadecuada |

Fig. 1.4. Ejemplos de escalas ordinales de respuesta

Refiriéndonos ahora al nivel de medición cualitativo o nominal, utilizamos este tipo de escala cuando aludimos a la presencia o ausencia de un atributo –síntoma presente o ausente, enuresis nocturna presente o ausente, uso o no uso del mecanismo de sublimación, afrontamiento exitoso del estrés o afrontamiento ineficaz, acierto o error en un test de desempeño– o al referirnos al tipo o clase específica de un atributo dado; por ejemplo, tipo de estructura –neurosis, psicosis o perversión–, tipo de neurosis –fóbica, histérica u obsesiva–, tipo de inteligencia predominante –fluida o



cristalizada–, estilo de afrontamiento más habitualmente utilizado –productivo/centrado en el problema, improductivo/centrado en el problema, productivo/evitativo o improductivo/evitativo–. En este tipo de escala, entonces, solamente se trata de establecer la pertenencia de un sujeto a una categoría en virtud de un atributo dado que la determina, sin implicar una jerarquía entre los individuos. El número utilizado para identificar cada categoría simplemente nombra o nomina –de ahí que se hable de nivel nominal de medición– la etiqueta de la categoría sin indicar un ordenamiento entre los integrantes de esa clase. Por ejemplo, el 1 puede reemplazar el nombre de la categoría neurosis, el 2 puede significar psicosis y el 3, perversión, pero estos números y el orden en que se presentan las categorías son intercambiables e, incluso, podrían elegirse otros, como 500, 1000 y 1500, sin que ello altere el sentido de la nomenclatura, ya que el número implica cualidad y no orden ni cantidad.

Resumiendo, los niveles de medición posibles en Psicometría van desde la simple nominación, cualificación o ubicación del individuo en una categoría (nivel nominal o cualitativo) hasta la cuantificación o determinación de la cantidad del atributo que el sujeto posee (niveles de intervalos y de cocientes), pasando por un nivel intermedio o semicuantitativo (ordinal) donde sólo se pretende establecer una jerarquía u orden entre las personas sin conocer la cantidad del atributo considerado. Cada nivel, desde el nominal hacia el de razones, incluye las propiedades del anterior y agrega otras nuevas. (ver Fig. 1.5.; en esta figura, las propiedades de cada tipo de escala que figuran en negritas corresponden a los que se agregan al pasar de un nivel de medición a otro).

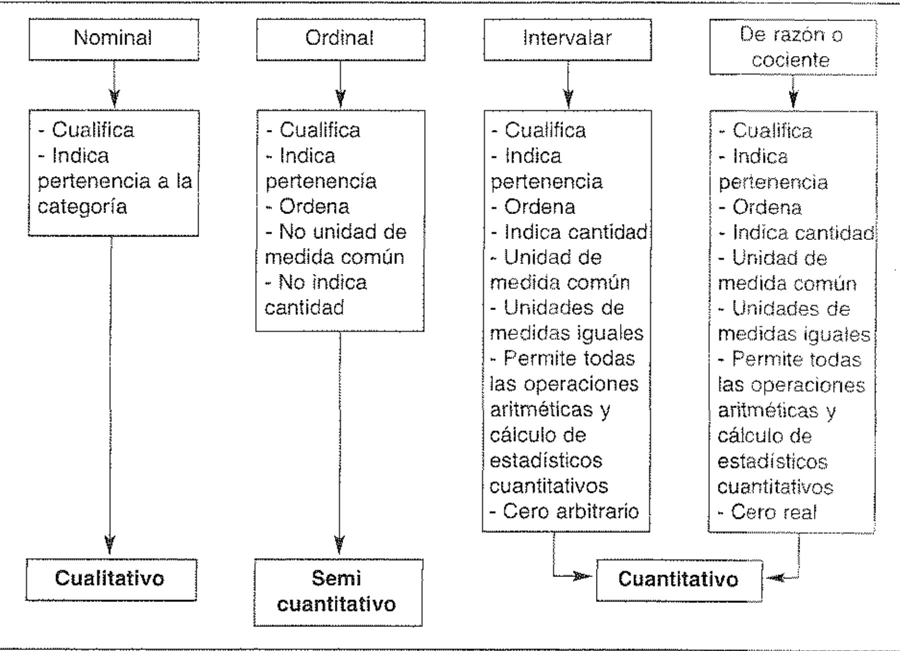



Fig. 1.5. Niveles de medición o tipos de escalas (Stevens, 1946)

El hecho de que trabajemos en uno u otro nivel de medición y, por lo tanto, de análisis, se relacionará con el interés particular del evaluador, con las características de la variable que se está estudiando y, finalmente, con el tipo de escala al que el instrumento empleado (test) es capaz de llegar. Así, en muchos casos –no en todos– es posible medir una misma variable en varios niveles de medición. Por ejemplo, la variable nivel educativo o escolaridad a la que ya nos hemos referido, puede ser medida en forma nominal o cualitativa (0: sin acceso a la educación formal; 1: con acceso a la educación formal), ordinal (1: sin educación formal o primaria incompleta; 2: primaria completa y secundaria incompleta; 3: secundaria completa y terciaria/universitaria incompleta; 4: terciaria/universitaria completa) o en un nivel de razones (años de educación formal cursados, que puede ir desde un cero real hasta la cantidad máxima posible). Resulta claro que la manera en que se diseñe cualquier categorización dada será arbitrariamente establecida por el evaluador o investigador, aunque esta decisión deberá fundamentarse en razones teóricas, técnicas o evolutivas, entre otras posibles, acompañándose de una precisa y clara división de las categorías que deben ser mutuamente excluyentes –un mismo individuo no puede ser ubicado en más de una categoría a la vez porque, de otra manera, la clasificación sería confusa e inútil–. La clasificación debe ser, entonces, exhaustiva –no debe dejar ningún individuo fuera de ella– y las categorías incluidas en ella, mutuamente excluyentes.

Una vez aclarado el punto que pretendíamos ilustrar, referido a que las técnicas psicométricas no deben necesariamente asociarse a la medición cuantitativa, sino que también pueden usarse en ellas escalas nominales u ordinales, volveremos a tratar el tema de los instrumentos psicométricos y proyectivos.

Es fundamental tener en cuenta que, tal como señalábamos antes, existen dentro del repertorio de pruebas disponibles, las *psicométricas* y las *proyectivas*. Teniendo en mente la importancia de integrar diferentes aspectos de las personas, siempre es deseable que, a la hora de efectuar procesos concretos de evaluación psicológica, se utilicen en forma conjunta herramientas provenientes de ambas vertientes. Sin embargo, este texto se dedicará solamente a desarrollar en profundidad cuestiones vinculadas a las primeras, sin perder de vista la necesidad de permanente integración entre los dos tipos de datos generados por ambas clases de herramientas en el ámbito de aplicación en el que el psicólogo trabaja diariamente.

Las personas poseen distintas características que, a juicio del evaluador, será más pertinente apreciar desde un punto de vista cualitativo, mientras que otras podrán ser evaluadas desde una medición cuantitativa o desde una valoración ordinal. En este punto, es importante pensar que en la realidad no existe tal división: cada individuo es una totalidad y su comportamiento también lo es; la diferenciación que hacemos de sus atributos o componentes es una distinción que parte de nuestra manera de analizar el fenómeno pero que no está en la realidad. Ya el mero hecho de hablar de "*atributos*" coloca al evaluador o investigador en una postura con respecto al tema; más aún, decidir cuantificar o medir cualitativamente u ordinalmente también es una decisión que depende del evaluador y no de la realidad donde se da el fenómeno en estudio. Es un problema de decisiones y de enfoque teórico el que se utilice uno u otro nivel de medición: los datos pueden manejarse, así, mediante categorías cuantitativas, ordinales o cualitativas e incluso, en otros casos, mediante métodos interpretativos, que son parte del terreno de las técnicas proyectivas. Merece resaltarse que en el ámbito de la Psicología aplicada, la tarea de evaluación integra modelos proyectivos y psicométricos puesto que los individuos no poseen un único nivel de análisis; mantener alguno de ellos sin abordar implica, sin dudas, dejar a un lado alguno de esos



aspectos, en una actitud reduccionista que restringe la acabada comprensión del fenómeno o problema en estudio (Bericat, 1998). De esta manera se integran en las baterías de instrumentos de evaluación las técnicas provenientes de ambas orientaciones, acompañadas, casi siempre, por la entrevista que, como decíamos en otro apartado, es el acercamiento más directo a la persona concreta que tenemos delante. La decisión última sobre qué pruebas se incluirán en esta batería es, como puede comprenderse, un problema técnico, metodológico, teórico y, en última instancia, epistemológico.

## 1.3 Los instrumentos psicométricos

En el apartado anterior mencionábamos que el psicólogo especialista en evaluación cuenta con un arsenal de instrumentos a su disposición, entre los que figuran las técnicas psicométricas, que son el objeto de interés para este texto. Para evitar efectuar un listado de varias definiciones posibles del término *instrumento psicométrico*, hemos elaborado una que sintetiza los aportes de los autores enumerados en el recuadro que sigue.

> Una ***técnica, prueba, test, escala o instrumento psicométrico*** (usándose cualquiera de estos términos como sinónimos posibles) se define como un dispositivo o procedimiento en el que se obtiene una muestra de *comportamiento de un examinado en un dominio específico, subsiguientemente evaluado y puntuado usando procedimientos estandarizados* (Anastasi & Urbina, 1998; Cohen & Swedlik, 2001; García Cueto, 1993; Hogan, 2004; Martínez Arias, 1995; Santisteban Requena, 1990).

La caracterización anterior hace referencia a un *procedimiento que se ha estandarizado*, es decir, que se ha tipificado de manera explícita, puntualizando específicamente un dispositivo o un método de trabajo: qué tipo de materiales deben utilizarse, qué consignas, qué ítems o estímulos, en qué tiempos de administración se trabajará, si habrá o no tiempo límite, de qué forma se entregará el material, qué actitud asumirá el examinador frente al sujeto y frente a las vicisitudes de su desempeño, qué criterios se emplearán para puntuar los resultados, entre otras cuestiones a ser previstas. Este conjunto de procedimientos se ha especificado y estandarizado al momento en que el test ha sido validado y han sido calculadas sus normas, si corresponde, y debe respetarse al pie de la letra cada vez que se administre a un examinado o a un grupo de examinados en el ámbito de aplicación.

Este respeto a la letra de tales prescripciones se basa en que cualquier cambio que se produzca impedirá la comparación de resultados o desempeños. Si se mantiene constante el modo de administración y puntuación, entonces también se podrá inferir que cualquier variación dada estará determinada por diferencias en los sujetos mismos –ya sean rasgos estables en ellos o estados transitorios– y no por alteraciones no previstas en el dispositivo de examen o de valoración del rendimiento.

En el caso de que alguna modificación se introduzca, entonces, ello debe tenerse siempre en cuenta puesto que las mismas podrían afectar las respuestas de manera indeseada o impensada. El monto e importancia de tales alteraciones

deben considerarse en cada caso particular, a la luz de decidir invalidar el procedimiento o de considerarlo pero con las salvedades discutidas. Para aclarar este punto, por ejemplo, puede pensarse en las condiciones en que se nos toman a las personas muestras de sangre en los laboratorios de análisis clínicos; ellas son extraídas en determinadas condiciones de ayuno, que están dadas por razones químicas y biológicas, pero también se respetan a rajatabla para que todos los pacientes guarden estas mismas condiciones y se puedan comparar los resultados obtenidos, captando debidamente las variaciones individuales. Si cada individuo mantiene una cantidad de horas de ayuno variable, entonces no se sabrá si las diferencias entre sus resultados son producto esas irregularidades o proceden de verdaderas distinciones en la composición de la sangre que se desea estudiar. Análogamente, cualquier cambio en las coordenadas de una evaluación psicológica puede implicar efectos no previstos o indeseados sobre los resultados, generados por esas alteraciones y no por diferencias reales en el desempeño o en las respuestas de los examinados.

Aquí resuena, seguramente, a los psicólogos la noción de *encuadre* que hemos heredado del psicoanálisis, que mantiene el mismo espíritu que lo que aquí señalamos. El encuadre consiste, dentro de un dispositivo analítico o psicoterapéutico, en el propósito explícito de volver constantes ciertas condiciones (por ejemplo, el lugar y duración de las sesiones, su frecuencia y costo, la regla de la asociación libre) para que el material surgido en estas sesiones sea resultado de contenidos estables o transitorios presentes en cada paciente o analizante y no provengan de alteraciones en las coordenadas del encuadre mismo (Siquier de Ocampo, García Arzeno & Grassano, 1987).

Así, volviendo a las técnicas de evaluación, no sólo los procedimientos de examen se tipifican y se hacen constantes, sino que también se realiza lo propio con las formas de puntuación y con las condiciones de interpretación de las respuestas brindadas por los examinados. Ahora bien, relacionado de alguna manera con este punto también es importante retomar aquello que decíamos al comienzo de este capítulo sobre qué sentimientos y emociones suelen experimentar las personas que se son evaluadas desde el punto de vista psicológico. Temor, ansiedad, curiosidad, indiferencia, aburrimiento, entre otros, pueden aparecer como respuesta a la situación de ser evaluado y de sentirse examinado. Sus respuestas y, en conjunto, todo el material brindado, se vinculan con temas personales, tanto de la propia historia como de modalidad o estilo individual. Los resultados, asimismo, pueden tener una incidencia muy grande en determinadas decisiones que afecten concretamente su vida cotidiana, a la vez que pueden ponerse en juego cuestiones relacionadas con su autoestima: *saber cómo le fue* produce un movimiento en todas estas variables. Así, en virtud de todas estas ansiedades, temores e implicancias personales, es fundamental que el evaluador, más allá de su estilo personal, intente permanentemente mantener un clima de trabajo cómodo, distendido, ameno y relajado, que promueva la colaboración, el interés y la motivación del examinado. Ello, por dos motivos: el primero se vincula con el hecho de que si el evaluado trabaja cómodamente, dará su mejor rendimiento y su más esmerado empeño en contestar de manera sincera, colaboradora, comprometida y con la mejor voluntad de trabajo; el segundo motivo tiene que ver con una actitud ética: dado que sabemos de todas las ansiedades y temores que la evaluación despierta, tenemos, además, la obligación profesional y humana de hacer que esta situación sea lo más fácil y relajada posible para la persona con la que estamos trabajando. En este sentido, surge como adecuado a las circunstancias, recordar el concepto de *rapport*, entendido como los esfuerzos puestos en juego por el evaluador para generar en



el evaluado una actitud general de cooperación, despertar su interés y motivación y estimularlo a responder a los tests y entrevistas de la mejor manera posible, según los objetivos planteados en cada uno de ellos (Anastasi & Urbina, 1998). Para ello, el examinador debe mantenerse atento desde el primer contacto, y a lo largo de todo el proceso de evaluación, a las características personales del evaluado: su edad, su sexo, sus variables de personalidad, su nivel educativo, su pertenencia a determinados grupos y subgrupos culturales, profesionales o religiosos, sus modalidades de contacto con las demás personas, sus limitaciones físicas si las tuviere, sus preferencias e intereses, sus habilidades, sus peculiaridades psíquicas y biológicas, entre otros muchos aspectos. En base a todo esto, debe reaccionar de manera rápida e intuitiva, poniendo en juego su capacidad empática y sus mejores herramientas para relacionarse con otros. Esta actitud de constante vigilancia sobre el estado de la relación que se da entre examinador y examinado es un trabajo muy refinado y sutil, que requiere una permanente corrección del rumbo a cada paso de la sesión de trabajo y en cada nueva sesión, puesto que en el intervalo entre cada encuentro pueden advertirse modificaciones en el sujeto por la influencia de factores diversos. Así, el trabajo con técnicas debe ir siempre acompañado del establecimiento de un buen rapport que facilite la tarea y que implique una actitud de respeto hacia el evaluado.

Volviendo ahora a la definición de test, recordemos que habíamos puntualizado en que se trata de un *dispositivo o procedimiento en el que se obtiene una muestra de comportamiento de un examinado en un dominio específico, subsiguientemente evaluado y puntuado usando procedimientos estandarizados*. Ya nos hemos referido al tema de la estandarización de procedimientos. Examinemos ahora la parte de la definición que contempla el hecho de que *este instrumento brinde al examinador una muestra de comportamiento en un dominio específico*. Ello significa que el sujeto recibe una consigna que le indica qué se espera de él, qué se busca que haga (responder preguntas, armar un rompecabezas, hacer un dibujo, memorizar palabras, entre otras conductas posibles), y al responder a dichas instrucciones, generará un conjunto de comportamientos que el evaluador registrará cuidadosamente (respuestas a las preguntas, el armado concreto de rompecabezas en un tiempo dado, el dibujo efectuado que queda registrado en un papel, etc).

Estos comportamientos son una muestra de un universo de comportamientos posibles en el individuo, que se han dado bajo determinadas condiciones y que pueden generalizarse a algunas otras situaciones, con cierto margen restringido de certeza y razonabilidad. A su vez, dichos comportamientos están circunscriptos a un dominio específico, ya que toda prueba acota o *"recorta"* sus consignas y, por lo tanto, las respuestas esperadas, a un dominio específico y bien definido, referido a una o más variables psicológicas, como pueden ser la ansiedad, la memoria de corto plazo, la atención, la hipocondría, la inteligencia, las fobias. Es decir que la consigna y los materiales elegidos restringirán el dominio al que se quiere circunscribir los comportamientos que se desea registrar y analizar.

Expresándolo más coloquialmente, las condiciones del test pretenden lograr que el examinado emita una serie de comportamientos reducidos a un área específica de todos los comportamientos que habitualmente genera (por ejemplo, restringidos al área de las relaciones interpersonales, al de las habilidades espaciales o, por caso, a la de los intereses vocacionales).

Todo esto implica que un test no evalúa *todo* el comportamiento de una persona, sino una muestra de todos sus comportamientos posibles, reducidos a un área específica o *dentro* de un área específica. Por supuesto que esta restricción estará dada por

dos razones: qué se quiere evaluar en concreto y desde qué marco teórico se hará o interpretará dicha evaluación. El modelo teórico también circunscribirá y definirá los comportamientos, pues toda teoría implica un recorte dado de la realidad.

Ahora bien, a esta definición básica de prueba agregaremos la condición de que deben determinarse y aportarse *evidencias empíricas sobre la validez y la confiabilidad* de los resultados arrojados por la misma: los autores de la técnica u otros investigadores deben diseñar y llevar a cabo estudios empíricos que den idea al usuario de dos cuestiones básicas: por un lado, si el test mide o evalúa aquello que dice medir –validez– (ver cap. 2), y por el otro, si aporta resultados o mediciones en las que se pueda confiar, con un error de medición predecible y determinable –confiabilidad– (ver capítulo 4, confiabilidad y error de medición). Así, entonces, el concepto de prueba piscométrica se completa de la siguiente forma:

> Una ***técnica, prueba, test, escala o instrumento psicométrico*** se define como un *dispositivo o* ***procedimiento estandarizado*** *en el que se obtiene una* ***muestra de comportamiento*** *de un examinado* ***en un dominio específico, subsiguientemente evaluado y puntuado usando procedimientos estandarizados****, y que cuenta* ***con evidencias empíricas sobre la validez y la confiabilidad de los resultados*** *que arroja* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999; Anastasi & Urbina, 1998; Cohen & Swerdlik, 2001; García Cueto, 1993; Hogan, 2004; Martínez Arias, 1995). Ver Fig. 1.6.

En este punto, se vuelve importante centrar la atención en las diferencias entre las técnicas psicométricas y otro tipo de instrumentos disponibles en el área de la evaluación psicológica, como son las técnicas proyectivas y las entrevistas diagnósticas.

Las *técnicas proyectivas* deben su nombre al supuesto de que se basan en el principio de la proyección. Este es uno de los mecanismos de defensa con el que contamos los seres humanos para hacer frente a la ansiedad, al estrés y a las situaciones conflictivas o críticas. Tal repertorio de recursos yoicos se va formando a lo largo del ciclo vital y nos ayuda a afrontar situaciones de desintegración que son vividas como amenazas para el yo (Bellak, 1992). La proyección supone exteriorizar, *"poner en el afuera"* o adjudicar a otros aquellos contenidos inconscientes y preconscientes que forman parte de nuestra personalidad profunda. Pero para que este mecanismo se dé, es necesario que exista una situación de poca estructuración, esto quiere decir, que mantenga cierta ambigüedad y que no se halle muy clara o estrictamente definida. Así, las consignas y los estímulos que disparan las asociaciones o respuestas comportamentales en las técnicas proyectivas poseen escasa estructuración y están pensados de manera tan amplia que, potencialmente, pueden propiciar un repertorio de respuestas que tiende a infinito (Anastasi & Urbina, 1998; Rapaport, 1978).

Por supuesto que existen grados diversos de estructuración tanto en los materiales estimulares como en la amplitud de las instrucciones; por ejemplo, el Test Proyectivo del Dibujo de una Persona (Machover, 1949) posee una consigna que pauta mínimamente la actividad, dado que solicita el dibujo de una persona sin mayores especificaciones o aclaraciones para luego pedir el dibujo de una persona del otro sexo. La técnica del H.T.P. –Test del Dibujo de Casa-Árbol-Persona: House-Tree-Person– (Buck, 1948, 1992), que insta a graficar esos elementos, plantean una situación más estructurada



que la anterior, pero menos que la que se propicia en el Dibujo Kinético de la Familia (Burns, 1982; Burns & Kaufman, 1970); este último indica efectuar el dibujo de *la propia familia haciendo algo.* De esta forma, no sólo se restringen más los elementos a ser reproducidos como en el H.T.P. sino que también se agrega la salvedad de la actividad, pautando aún más la producción esperada.

Siguiendo esta línea de pensamiento, en un sentido muy general, podría considerarse que la entrevista libre también sea una especie de técnica proyectiva puesto que se asemeja a la página en blanco de los dibujos, en la que se *proyectan* los propios contenidos inconscientes a la manera de una pantalla cinematográfica o de un proyector multimedia.

Análogamente, examinando el gradiente de estructuración que presentan las técnicas proyectivas con estímulos gráficos y respuestas verbales, podemos citar el Test de Rorschach (Rorschach, 1921/1942); en él la consigna pide que se examinen las manchas de tinta y se diga qué se ve en ellas, advirtiendo al sujeto que no hay respuestas correctas o incorrectas y que distintas personas perciben diferentes cosas allí. El Test de Relaciones Objetales (Phillipson, 1983), en cambio, plantea una situación en la que aparecen tres clases de dibujos que representan situaciones de soledad, escenas de a dos y de tercero excluido con diferentes grados de estructuración-ambigüedad; se pretende que el examinado diga, a partir de las mismas, cómo se originó la situación, qué está sucediendo en ese momento y cómo termina, acotando aún más las respuestas posibles de lo que el Rorschach lo hace.

Los ejemplos anteriores permiten apreciar cómo todas las técnicas proyectivas comparten la característica de poseer poca estructuración en estímulos y consignas, dejando bastante abiertas las respuestas posibles que, potencialmente, son infinitas. También comparten el hecho de que todas ellas tienen un marco teórico en común, que es el psicoanálisis; algunas basan sus interpretaciones en los postulados freudianos, otras, en los kleinianos o postkleinianos; otras, recurren a Ana Freud. En algunas proyectivas, en particular, será factible efectuar diferentes interpretaciones según el autor que se tome como fundamento, pero el elemento en común es el marco de referencia psicoanalítico, evaluándose elementos y contenidos relativos a la personalidad profunda (Avila Espada, 1997; Hammer, 1957). Debe destacarse, sin embargo, que han existido algunos esfuerzos aislados para adaptar estos instrumentos a las interpretaciones derivadas de otros modelos, como por ejemplo las teorías perceptuales de la personalidad (Exner, 1995; Lindzey, 1961), sin embargo, el marco que las sustenta es, básicamente, el psicoanálisis.

Otra de las diferencias que las proyectivas y las psicométricas mantienen es que las primeras proponen la evaluación de la personalidad como un todo, en tanto que las segundas aíslan atributos diversos, valorándolos de a uno a la vez. Las proyectivas, si bien admiten separar y analizar individualmente ciertos componentes de la personalidad, tratan a los mismos como parte de un todo inter-relacionado y separable sólo a los fines de su estudio y análisis (Anastasi & Urbina, 1998). Si bien es cierto que, de manera general, las proyectivas brindan una visión más global de la personalidad del sujeto, debe tenerse presente que cada vez que se aíslan elementos –ya sea por medio de las técnicas proyectivas o por el uso de las psicométricas– esta disección siempre se efectúa de manera artefactual y a los fines de lograr una mejor comprensión de los mismos; pero los componentes de la personalidad y, más ampliamente, los atributos psíquicos de los seres humanos no son distinguibles en el comportamiento observable; su separación es, en todos los casos, un artificio para entenderlos mejor que resulta posible gracias al auxilio de un modelo teórico dado.

A diferencia de las proyectivas, los instrumentos psicométricos cuentan con consignas, estímulos y alternativas de respuesta altamente estructurados, a la vez que pueden fundamentarse en diversos marcos teóricos, entre los que el psicoanálisis es solamente una de las alternativas posibles. Ello sucede porque las variables evaluadas no corresponden a la personalidad profunda sino que más bien se trata de elementos tales como la inteligencia, las habilidades, la ansiedad, las dimensiones de la personalidad conceptualizadas según Kraepelin o Millon, por caso, el autoconcepto, la maduración visomotriz y conceptual o la memoria de corto plazo. En este sentido, y como no se sustentan en mecanismos proyectivos, se evita deliberadamente plantear situaciones poco estructuradas para acotar el rango de respuestas posibles y facilitar la puntuación y valoración de las mismas según criterios estrictos y, en lo posible, unívocos. En auxilio de este propósito se trabaja con materiales e instrucciones fuertemente estructurados, así como con opciones de respuesta preestablecidas y/o con criterios de puntuación minuciosamente definidos. Desde este punto de vista, algunos autores consideran a la entrevista dirigida o cerrada –que consta exclusivamente de preguntas preestablecidas– como una técnica psicométrica, debido a su alta estructuración; en cambio, la entrevista libre, en la que el sujeto habla sin ninguna restricción sobre aquello que desee y del modo en que lo desee, puede ser categorizada como una técnica proyectiva. De todas maneras, cabe destacar que otros teóricos de la evaluación ubican a las entrevistas, cualquiera sea su tipo o grado de estructuración, en una clase aparte de las técnicas psicométricas y las proyectivas. Más allá de estas diferenciaciones, es importante destacar que la entrevista es, en la inmensa mayoría de las situaciones de evaluación, una herramienta indispensable de acercamiento al examinado, puesto que en ella se propicia una interacción tan directa que permite acceder fácilmente a la problemática o características del mismo, así como despejar dudas, aclarar puntos oscuros y corroborar o refutar diferentes hipótesis interpretativas que pudieran haberse generado a partir de los resultados arrojados por las técnicas (Albajari, 1996). A la vez, la situación de entrevista en sí misma resulta familiar y cotidiana para la mayoría de las personas, mucho más que la administración de herramientas psicométricas o proyectivas, que no se asemejan a situaciones habituales en la vida de los individuos comunes.

Tales son, básicamente, las diferencias formales más inmediatas que pueden establecerse entre ambos grupos de instrumentos, psicométricos y proyectivos. Ahora bien, es importante comprender que en la tarea de evaluación, en la inmensa mayoría de los casos, se trabaja integrando la información que deriva de ambas fuentes, atendiendo a ambas clases de aspectos de los seres humanos: los de la personalidad profunda y los referidos a otras variables que no se encuentren en esta esfera. Más allá de la adhesión que pueda tener el profesional hacia los postulados psicoanalíticos, algunos atributos personales escapan a los métodos de evaluación con que se trabaja en las técnicas proyectivas, así como otros escapan a la evaluación psicométrica, dada la diferencia en la naturaleza de tales características, y dadas las distinciones planteadas en los distintos abordajes teóricos y técnicos que ambos tipos de aspectos implican. La riqueza de una labor profesional bien realizada radica, la mayor parte de las veces, en la integración aceitada de ambas clases de datos.



## 1.4 Los test como operacionalizaciones de constructos teóricos

Enfocando ahora el tema de las técnicas psicométricas desde el abordaje característico de la Metodología de la Investigación, es necesario puntualizar que todo test se basa en un modelo dado, como ya se especificó párrafos atrás. Asimismo, ese marco teórico que fundamenta la técnica apela a diversos conceptos o constructos –construcciones ideales, formuladas por la mente humana para explicar determinados aspectos o fenómenos de la realidad -; por ejemplo, la noción de ansiedad es un concepto construido por algunos autores, que no existía de manera natural en la realidad. Veamos cómo opera un teórico o un investigador: observa determinados eventos de la realidad– tales como la sudoración o la palidez de una persona, el temblor de su voz o de sus manos, el hecho de que comente hallarse muy preocupado o *"nervioso"*, que relate que no puede dormir en las noches, que se verifique taquicardia al tomarle el pulso, que se lo vea caminar de un lado a otro demostrando inquietud, etc. -, y adjudica cierta unidad a ese conjunto de fenómenos observados, que conceptualiza bajo el rótulo de *ansiedad*. Así, esas manifestaciones orgánicas o fisiológicas (ej., sudoración), motrices (ej., su caminar inquieto) y cognitivas (ej., preocuparse) serían indicadores de un concepto teórico y abstracto que las agrupa y de un fenómeno que las contiene: la ansiedad. La ansiedad, en sí misma, no es observable, perceptible ni pasible de ser medida; como concepto, no tiene existencia real sino ideal; sólo existe en la mente de quien la ha descripto y de quien la estudia o la entiende al leer en un texto. Lo que sí tiene existencia real y es apreciable mediante los sentidos, observable o medible, son sus manifestaciones o indicadores: los signos y síntomas de la ansiedad, es decir aquellos indicios que se pueden observar, oír o tocar en forma directa y aquellos de los que el sujeto que los experimenta puede informarnos.

Así, una vez que se ha definido y descripto este concepto desde un modelo teórico, apelando, a su vez, a otros conceptos de esa misma teoría que ayuden a caracterizarlo, se procede a *operacionalizarlo; operacionalizar un concepto o definirlo operacionalmente implica "bajar" su definición abstracta a la empiria mediante la identificación de indicadores observables que den cuenta de la ocurrencia de este fenómeno en la realidad*. De esta forma, todo constructo teórico implicaría una definición teórica o conceptual y una operacional. La teórica se ocupará de examinar distinciones relevantes a la luz del modelo desde el que se define el concepto, en interjuego con otros conceptos pertenecientes a esa misma teoría; por otra parte, la definición operacional u operacionalización implicará elaborar un listado de indicadores u observables empíricos que den cuenta de la presencia o ausencia de dicho fenómeno en la realidad, o bien del grado en que se registra su ocurrencia.

Mediante esa prueba empírica la teoría se corrobora o no con datos de la realidad, confirmándose o debiendo reformularse a la luz de esos datos reales. Este es el camino que va de la teoría a la realidad y de la realidad a la teoría y que es la base de toda tarea de investigación científica. Ahora bien, siguiendo este razonamiento podemos entender los instrumentos psicométricos como un conjunto de indicadores de un concepto o constructo teórico; ese conjunto de indicadores observables o medibles son, ni más ni menos que los *ítems, elementos o reactivos* del test. Ese test será elaborado en conformidad y coherencia con ese marco de referencia teórico que le da sustento y, por ende, los ítems que lo forman serán indicadores que se habrán operacionalizado también en forma consistente con esa teoría. Una vez derivados, entonces, los indicadores o reactivos, los resultados obtenidos a partir de las respuestas dadas a

ellos por parte de personas reales y concretas, permitirán volver a la teoría para examinarla otra vez, a la luz de esos resultados empíricos (Fig. 1.6.).

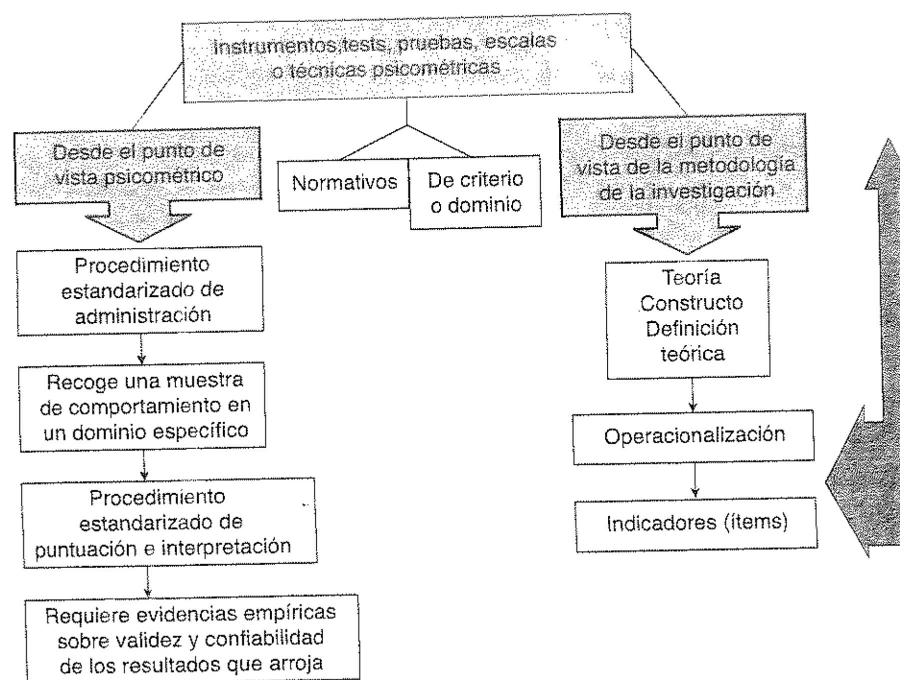



Fig. 1.6. Caracterización de los instrumentos psicométricos según la Psicometría y la Metodología de la Investigación, y según el uso de normas o criterios.

Ya que nos hemos referido a los ***ítems, elementos o reactivos de un test***, diremos que éstos son la *mínima unidad distinguible en él, consistentes en cada una de las pequeñas tareas o actividades que el individuo debe realizar para responder a la consigna*. Estas actividades pueden consistir en efectuar un dibujo, encajar piezas, elegir la opción correcta, la opción preferida o la más frecuente, copiar formas, etc. El total de los ítems forma la escala o prueba, pero en algunos casos, cuando existe diversidad de tareas o de contenidos porque dentro de la variable evaluada es posible distinguir dimensiones subyacentes, variables latentes o factores, entonces puede hablarse de *subtests*, y aún de *subescalas*. En este caso, varios ítems forman un subtest o *subprueba* y varios subtests integran una *subescala*. Dos o más subescalas forman la escala total o instrumento psicométrico. Sin embargo, la alternativa más simple consiste en una escala unidimensional –una sola dimensión– compuesta por un grupo homogéneo de ítems que no se agrupan según el tipo de tarea en subtests ni en subescalas.



## 1.5 La noción de escalamiento

Recapitulando a partir de los puntos desarrollados en el apartado anterior, desde un doble criterio para definir los instrumentos psicométricos según una perspectiva técnico-psicométrica y metodológica, podemos concluir que una prueba psicométrica supone:

Cuadro 1. Conceptos implicados en la noción de instrumento psicométrico

- **Una colección de indicadores relativos a un dominio de comportamiento precisamente definido**
- Implica:
  - La **medición de un rasgo** o atributo de un sujeto que se ha operacionalizado desde un modelo teórico
  - La noción de **escalamiento**
  - Capacidad para **discriminar** diferencias individuales entre las personas
  - **Validez y confiabilidad** verificadas empíricamente (la noción de confiabilidad incluye, entre otras notas, cierta **estabilidad temporal)**
  - Determinados **atributos formales**
  - Estar enmarcado dentro del **área** de la **Evaluación Psicológica**

Como puede fácilmente advertirse, el Cuadro 1 se refiere, en primer término, a las **herramientas psicométricas** examinadas desde el *punto de vista metodológico*, entendidas como un **conjunto de indicadores** u observables empíricos de un concepto o constructo teórico, tal como se comentaba algunos párrafos atrás. Ahora bien, esos indicadores se han restringido previamente, a un dominio específico y acotado de comportamiento, dado que no se puede medir u observar *todo* el comportamiento de un individuo. Así, las consignas, los ítems y los materiales que sirvan de estímulo estarán dirigidos a circunscribir los comportamientos que se espera desencadenar a un área o dominio específico del comportamiento global del sujeto (por ejemplo, lograr que el examinado se autodescriba en relación a sus intereses vocacionales, que resuelva un problema de razonamiento mecánico o que opine sobre su conformidad acerca de las condiciones generales de su vida cotidiana).

Desde el *enfoque psicométrico* supone la **medición de un rasgo**, característica o atributo de una persona, que estará **operacionalizado de manera coherente con un marco teórico** que se ha tomado como fundamento y que, por supuesto, engloba al constructo que se pretende evaluar con el instrumento del que se trate.

Se halla involucrada, además, la noción de **escalamiento**, que significa, ni más ni menos que la posibilidad de *convertir o traducir las respuestas brindadas por los sujetos a una puntuación*. Como hemos visto en el apartado 1.2. de este capítulo, estas puntuaciones no necesariamente deben ser cuantitativas cuestión que podría sorprendernos, ya que las palabras *puntuación* y *psico-metría* suelen estar asociadas en el lenguaje común, a la cuantificación; pero si recordamos lo leído en relación a escalas o niveles de medición, esta salvedad debería quedarnos clara. Sigamos, ahora, con la noción de escalamiento.

Las respuestas brindadas por un sujeto (sus aciertos o fallos en un test de rendimiento, las preferencias que manifestó sobre cómo encarar las relaciones interpersonales o sus contestaciones a un listado de síntomas que pueden estar aquejándolo),

que se codificarán, como antes comentábamos, en una forma preestablecida y de manera pautada o altamente estructurada, se agruparán para ser resumidas en una puntuación final que brindará una idea general sobre el conjunto de las respuestas. Por ejemplo, en el Test de Matrices Progresivas de Raven (Raven, Court & Raven, 1993), que mide la capacidad eductiva, el sujeto es enfrentado a 60 ítems que debe resolver, eligiendo la respuesta correcta de entre seis u ocho opciones posibles. Esas respuestas quedarán registradas en lo que llamaremos ***protocolo***, que es simplemente eso: un *registro de las respuestas brindadas por el sujeto*. Más tarde el examinador codificará esas contestaciones según el sistema tipificado en el manual (recordar las consideraciones que hacíamos antes sobre la necesidad de respetar a la letra estos procedimientos para poder comparar resultados) para, finalmente, obtener una puntuación global que, desde el punto de vista estadístico y metodológico, es un *índice*, es decir una *puntuación construida* que nos servirá para resumir la serie de respuestas dadas por el sujeto. Asimismo, la medición posible del constructo evaluado por la técnica debe tener cierta ***estabilidad temporal***, que puede ser elevada o baja, pero debe existir un mínimo de ella para que la evaluación tenga algún sentido (ver cap. 4). Si bien es cierto que hay variables más inestables que otras. Por ejemplo, los estados emocionales transitorios son más inestables que los niveles de maduración conceptual a los que acceden los niños; estos últimos, a su vez, son más inestables que los rasgos de la personalidad adulta. Los estados emocionales pueden variar en cuestión de minutos o segundos pero tienen alguna duración mínima en el tiempo que justifica el esfuerzo de medirlos; los niveles de maduración alcanzados en la niñez, si bien se mantienen inalterados poco tiempo, registran una duración algo mayor, en tanto que los rasgos personológicos adultos tienden a guardar una estabilidad mucho más pronunciada a lo largo de esa etapa del ciclo vital.

Otro problema que debe atenderse es que el instrumento sea realmente capaz de captar las diferencias individuales que existen entre las distintas personas en cuanto al rasgo o atributo medido, es decir, de **discriminar**. Aquí debemos hacer una importante aclaración, en el sentido de que en nuestro medio, frecuentemente, se usa este vocablo con connotaciones negativas, puesto que se lo identifica con el hecho de no permitir el acceso de una persona a un lugar o posición determinados por razones que no están relacionadas con sus aptitudes, su capacitación o, simplemente, sus derechos civiles (por ejemplo, no admitir a un joven en un local bailable porque su apariencia no corresponde al estereotipo clásico de una clase social, o no seleccionar a alguien capacitado para un empleo dado porque posee una discapacidad que no lo limita para la adecuada ejecución de sus funciones).

En psicometría, el término discriminación no cobra este sentido, sino el de diferenciar en forma adecuada entre sujetos que poseen el atributo analizado en gran medida respecto de quienes lo poseen en un grado más bajo o directamente no lo poseen. Para que se comprenda mejor mediante el empleo de un ejemplo más cotidiano, cuando tomamos un examen a los alumnos, decimos que una pregunta del mismo no ha discriminado cuando aquella no sirvió para detectar quiénes realmente sabían el tema, diferenciándolos de aquellos que no lo habían aprendido. Esto sucede cuando incluimos una pregunta tan amplia o tan fácil de responder que casi cualquier respuesta de sentido común resulta adecuada, de manera que todos la contestan exitosamente, sin necesidad de poner en juego los contenidos específicos que se intentaba evaluar y, por lo tanto sin que el docente que corrige estos exámenes pueda diferenciar o *discriminar* si el alumno ha aprendido o no ese concepto.



Este es el sentido de la palabra *discriminación en psicometría: el instrumento debe servir para captar diferencias individuales, puesto que ese es su sentido: darnos información sobre las características únicas de un sujeto, aquellas que lo diferencian de los demás* (García Cueto, 1993; Martínez Arias, 1995).

El test debe tener, por otra parte, y tal como también indicábamos en el Cuadro 1, **estudios empíricos** que hayan arrojado **evidencias sobre la validez y la confiabilidad** de las puntuaciones que éste permite obtener (en los caps. 2 y 4 se desarrollarán estos conceptos con mayor detalle). Por el momento nos limitaremos a afirmar que los *estudios de validez* nos permiten conocer de qué manera el autor del test u otros investigadores aportaron *evidencias empíricas de que este instrumento está evaluando el constructo que promete evaluar* o está midiendo otra/s variable/s (ver cap. 4). Sencillamente expresado: ¿cómo podemos estar seguros de que una prueba que se ha denominado, por ejemplo, Escala de Inteligencia de xx está realmente midiendo la inteligencia de las personas, o sencillamente está poniendo a prueba su memoria o cualquier otra capacidad? El autor original u otros investigadores deben aportar pruebas de que lo que se promete medir, efectivamente, se está midiendo. Las dudas sobre este punto surgen porque en Psicología los conceptos con los que trabajamos poseen una múltiple determinación y, además, no son observables, tal como comentábamos en párrafos precedentes. La ansiedad, la neurosis, la inteligencia, las actitudes o los prejuicios no se pueden observar directamente sino sólo a través de determinados indicadores apreciables empíricamente (en este caso, los ítems del test). El investigador debe demostrar que esa prueba mide ese constructo según la teoría que se ha propuesto como fundamento, y con indicadores (ítems) adecuados para efectuar una medición ajustada.

Por otra parte, la *confiabilidad de las puntuaciones obtenidas mediante el test* alude a la *confianza que podemos tener acerca de los resultados concretos que esta prueba brinda* al medir el constructo aludido. Este concepto es muy complejo e involucra varios aspectos en su definición, tales como la estabilidad temporal, la consistencia de sucesivas mediciones, la consistencia interna de los ítems del instrumento, los errores que se producen al medir y, también, la homogeneidad de los reactivos al interior de la técnica. Todas estas cuestiones se examinarán en detalle en el cap. 4.

El test, por otra parte, contará con ciertas **propiedades o características formales,** los materiales empleados, la forma de administración, el uso de tiempo límite o no, entre otros aspectos a ser considerados. En líneas generales, podría decirse que es factible realizar una multiplicidad de clasificaciones de las pruebas según sus aspectos formales, que variarán notablemente en base al criterio que se adopte para efectuar esa categorización. Sin embargo, las más comúnmente efectuadas se refieren al objetivo de la pesquisa, a los materiales empleados, a la forma de administración, al objeto de la evaluación, al tipo de respuesta, al formato, al uso del tiempo y, fundamentalmente, a la base sobre la que se comparan las respuestas o el desempeño del sujeto evaluado.

Refiriéndonos ahora al ***objetivo de la pesquisa***, puede hablarse de *tests de diagnóstico* y de *tests de screening*. Los primeros intentan brindar al usuario una evaluación detallada y pormenorizada de una situación, un atributo o un estado o rasgo dado, mientras que los tests de *screening, rastrillaje* o *despistaje* se usan para detectar *riesgo*; es decir que dan una evaluación poco detallada, preliminar y que debe profundizarse. Por ejemplo, si se desea detectar personas en riesgo de padecer trastornos de alimentación en una población amplia, no es posible aplicar a tanta gente una batería completa de instrumentos en virtud de los tiempos, el esfuerzo y los costos

involucrados. Los tests de despistaje y screening son breves y altamente sensibles; es más, es importante que sean *sensibles en demasía y poco específicos*, ya que deben reaccionar fácilmente ante los mínimos indicadores de riesgo potencial. Cuando se trabaja de esta manera – tal como suele hacerse en la investigación epidemiológica – es preferible que se detecten casos *falsos positivos*, en los que luego, en una evaluación más profunda, no se compruebe riesgo o patología, antes que se deje sin identificar casos verdaderamente positivos que, de otra manera, pasarían desapercibidos, en virtud de la baja sensibilidad del test. Una vez localizados estos casos mediante el instrumento de despistaje, se vuelve a examinar a esos sujetos pero en esta oportunidad con técnicas de diagnóstico, que brindarán una información más acabada y completa, permitiendo confirmar los resultados iniciales o descartarlos. Estos instrumentos de screening funcionan como un filtro o cedazo que separa los elementos más *gruesos* (indicadores poco específicos), para que sean examinados en detalle mediante técnicas diagnósticas, decidiendo a posteriori si existían razones de peso para seleccionarlos o si, simplemente, se trató de un caso falso positivo. Para ejemplificar, teniendo en cuenta otras disciplinas, podemos tomar como ejemplo las campañas que anualmente se organizan para la detección de personas con diabetes. Se invita a la población a acercarse durante un período dado a hospitales y puestos sanitarios y se le pide que, habiendo cumplido ciertas condiciones de ayuno o alimentación, permitan que se les extraiga una gota de sangre que cae en una tira reactiva colocada en una aparato electrónico, que reacciona ante la cantidad de glucosa presente en esa sangre, *traduciendo* esa cantidad a un valor numérico que aparece en la pantalla. Si ese valor se encuentra por debajo de cierto límite establecido, el sujeto se va a casa tranquilamente, en tanto que si sobrepasa esa puntuación de corte se lo instará a que consulte con un médico que realizará otros controles (exámenes de laboratorio tradicionales, curva de glucemia, etc) para determinar si se trató de una elevación circunstancial por efectos del estrés, de una medicación o de una ingesta alimentaria atípica, entre otras causas posibles. Si los resultados no se confirman y se dirime que algún factor extraño produjo esa medición elevada pero atípica, se trató de un caso falso positivo; en cambio si las pruebas ulteriores de diagnóstico confirman la información dada por el test de riesgo, entonces se tratará de un caso positivo de diabetes. Es, entonces, fácil comprender por qué los instrumentos de screening son de administración y evaluación breve y sencilla: porque se trabaja con un gran número de personas en un corto lapso.

Las *técnicas de diagnóstico*, en cambio, nos darán mucha más información, más profunda y pormenorizada, sobre la variable que se esté evaluando; sus tiempos de administración y evaluación serán, por ende, mayores, pero permitirán arribar a una descripción más acabada, complementada por una cierta cantidad de datos y detalles. De manera inversa a las técnicas de screening, serán menos sensibles, pero muy específicas.

En cuanto a los ***materiales y medios empleados***, se encuentran en el mercado *tests de lápiz y papel, de materiales manipulables, de estímulo oral y respuesta oral, de estímulo gráfico y respuesta oral, y de estímulo oral y respuesta escrita*, entre otras múltiples posibilidades disponibles.

En relación con la ***forma de administración***, puede decirse que existen dos grandes grupos de instrumentos: los de administración individual y los autoadministrables. Los primeros requieren, por el tipo de tarea implicada, la interacción personalizada de un examinador con un examinado, en tanto que los autoadminstrables están especialmente diseñados para que sus consignas, sus materiales y sus ítems sean tan



claros que el sujeto sea capaz de dar respuesta a ellos sin ayuda, o con una mínima guía del evaluador; así, este tipo de modalidad permite, indistintamente, que la técnica sea administrada en forma individual (a un sujeto) o colectiva (a varias personas en un mismo momento). De esta forma, se brinda la posibilidad de acortar tiempos y costos cuando deben realizarse evaluaciones grupales y supernumerarias.

Con respecto al ***objeto de la evaluación***, suele hablarse de tests de habilidades, de potencia, de personalidad, entre otros. Básicamente esta categorización alude al *gran grupo* de conceptos o variables al que pertenece aquel constructo que se quiere evaluar.

Tomando ahora como criterio clasificatorio el ***tipo de respuesta***, las pruebas psicométricas pueden ser de formato dicotómico, likert, de diferencial semántico, de opción múltiple, de resolución de tareas específicas, de valoración de éxito o error, entre otras formas posibles (Cohen & Swerdlik, 2001).

La *respuesta dicotómica* es la que pone al examinado ante la obligatoriedad de decidir su respuesta entre dos opciones polares (generalmente verdadero-falso o si no); de esta forma, debe suspender pensamientos tales como *"depende de la situación"* y forzarse a sí mismo a escoger.

La *respuesta likert* implica un ordenamiento de las opciones según un gradiente que va desde la máxima aceptación al máximo rechazo, o viceversa. También pueden plantearse, por ejemplo, respuestas likert en términos de frecuencia de aparición temporal de los comportamientos u otras alternativas de categorización posibles (Cortada de Kohan, 2004). La cantidad de posiciones que tiene una respuesta likert estará dada por determinados propósitos que tenga la técnica así como por las características de la población a la que ella se destina. Así, pueden emplearse respuestas likert de tres o de cinco posiciones según se quieran diversificar o no las respuestas posibles; se utilizarán likerts de cuatro puntos cuando se quiera impedir que los examinados se vuelquen mayoritariamente a la respuesta neutra, ubicada en el medio de las escalas cuando son impares. Las likerts de siete o nueve posiciones suelen evitarse en nuestro medio en virtud de que tantas posibilidades suelen confundir mucho a los sujetos, ya que algunas de ellas pueden superponerse según la apreciación de cada evaluado. De manera general, el nivel educativo de los examinados, así como el de comprensión lectora en particular, determina en gran parte la conveniencia de utilizar cada una de las alternativas de respuesta likert antes detalladas.

El *diferencial semántico* (Osgood, Suci & Tannenbaum, 1957) es una forma de respuesta que prevé una escala, generalmente de siete o de nueve puntos, en cuyos extremos se ubican dos adjetivos o expresiones con significados contrapuestos, y se solicita al examinado que marque en qué punto de ese continuo ubica su parecer con respecto a aquella variable que se está evaluando, en términos de valorar significados posibles de ser atribuidos. Los puntos intermedios corresponden a hitos de una escala graduada que indica cierta distancia respecto de los polos del diferencial en los que se colocan los adjetivos polares.

En los tests de *resolución de tareas específicas* se pide al sujeto que realice una producción determinada (copiar tarjetas, dibujar bajo determinadas condiciones, formar figuras con cubos, resolver un laberinto o un rompecabezas, codificar números arábigos en una simbología especial que debe ser aprendida, recordar palabras y repetirlas, etc). Si bien en estos casos las respuestas quedan registradas ya sea en el propio dibujo o producción realizados por el evaluado o porque han sido consignadas por el evaluador mediante pautas dadas, suelen, más tarde, re-codificarse según criterios diversos; por ejemplo, como éxito-fallo, o como respuesta completamente adecuada –parcialmente adecuada– inadecuada, entre otras variaciones posibles.

/

Según su ***formato***, las técnicas pueden dividirse en inventarios, cuestionarios, escalas clásicas, encuestas, protocolos de entrevistas dirigidas y protocolos de observación, entre las categorías más representativas. Los *inventarios* son *listados de afirmaciones* que el sujeto debe leer y responder, por ejemplo, según opciones de verdadero-falso o según un gradiente en el que expresa su acuerdo o desacuerdo (likert). Los *cuestionarios*, en cambio, son *listados de preguntas* que el sujeto debe responder según un formato preestablecido (sí–no, o según grados de conformidad o frecuencia con respecto a lo que se está interrogando). Es importante tener en cuenta que algunos autores no distinguen entre cuestionario e inventario, nombrando indistintamente a ambos tipos de técnicas, más allá de que estén formados por afirmaciones o por preguntas (Anastasi & Urbina, 1998).

Las *escalas*, por su parte, suelen identificarse con algunos tests de rendimiento que tienen un formato diferente de los inventarios o cuestionarios, y que se componen, por ejemplo, de tareas que el sujeto debe resolver o de preguntas que debe contestar para reflejar algún conocimiento o destreza. Sin embargo, en un sentido amplio, *todos los instrumentos psicométricos son escalas dado que implican*, como decíamos antes, *la noción de escalamiento*, que significa *convertir las respuestas de los examinados a una puntuación que las resume* (Martínez Arias, 1995).

Las *encuestas* suelen tener un formato similar al de los inventarios y se utilizan para recolectar opiniones o actitudes de las personas sobre algún tema en especial, sobre un servicio o sobre costumbres y preferencias. El estilo de respuesta es generalmente similar al de los cuestionarios e inventarios y también sucede que ciertos autores no las diferencian de aquéllos.

Las *entrevistas dirigidas* son listados de preguntas preestablecidas que se hacen oralmente al sujeto en una administración individual y que deben responderse oralmente, con la posibilidad de ampliar esas contestaciones por medio de comentarios y detalles. Para la codificación de las respuestas se utiliza un protocolo preimpreso que permite un sencillo y breve registro de aquéllas.

Finalmente, las *hojas de registro o protocolos de observación* son también formularios preimpresos que establecen qué aspectos específicos deben observarse en determinados comportamientos o interacciones, en los que el examinador u observador va codificando en una forma abreviada y rápida aquellos atributos que son objeto de su evaluación (Casullo, Figueroa & Aszkenazi, 1991).

Como se puede apreciar, la clasificación de los instrumentos psicométricos según su formato se vuelve algo complicada en tanto las categorías que se utilizan no siempre son mutuamente excluyentes. Sirvan los párrafos anteriores como ejemplo de ello, ya que la idea es transmitir un panorama general sin bucear demasiado en sutilezas técnicas que no sean de interés para el usuario.

En cuanto al ***uso del tiempo***, existen *técnicas que no fijan un límite* temporal para finalizar la tarea, sino que permiten que el examinado trabaje libremente y a su ritmo. Otras, en cambio, establecen un *límite preciso* luego del cual se suspende la tarea, llegando hasta el punto al que se haya arribado en la actividad. Esta diferencia se decide en base a qué es lo que se quiere evaluar y cómo se desea hacerlo. Finalmente, otros tests permiten que se trabaje libremente pero *toman nota del tiempo* para valorarlo en una forma determinada, como por ejemplo, premiando al sujeto con una mayor puntuación por efectuar la tarea más rápidamente o usando los tiempos muy lentos o muy rápidos como información relevante para efectuar determinadas interpretaciones o elaborar algunas hipótesis.



Ahora enfocándonos en la ***base sobre la que se valoran o comparan las respuestas o desempeño del sujeto evaluado***, podemos clasificar los instrumentos psicométricos en *tests normativos* y *tests de criterio o de dominio* (Cohen & Swerdlik, 2001; Martínez Arias, 1995). (Fig.1.6.).

La primera modalidad de valoración de los resultados corresponde a las técnicas psicométicas que se rigen por *baremos o normas estadísticas*, llamados tests normativos. Ellas comparan el rendimiento o respuestas de una persona individual con el rendimiento promedio registrado por una muestra normativa o de tipificación, es decir, por un grupo de individuos homogéneos al examinado, según edad, sexo, hábitat y otras condiciones que pudieran afectar a la variable evaluada. Así, el baremo o norma estadística es un cuadro de doble entrada en el que se consignan el promedio de las puntuaciones obtenidas por esa muestra de sujetos y su dispersión o desviación típica (es decir, su distancia relativa respecto de la media o promedio). Este baremo estará dividido según sexo, edad, hábitat u otras variables posibles, de manera que figurará el rendimiento medio y su correspondiente desvío para cada intervalo de edad, para cada sexo y según el lugar de residencia del sujeto. Entonces, para la evaluación de un caso individual se valorará el puntaje obtenido por el examinado contra la puntuación media de la muestra de sujetos de su misma edad, sexo y hábitat, considerando también la amplitud del desvío, con el objetivo de poder establecer si el desempeño de nuestro examinado se ubica en la franja media, por encima de ese rendimiento, o por debajo. Expresándolo en forma sencilla, el baremo permite que el examinador valore el desempeño de un sujeto a la luz del desempeño promedio observado por sujetos semejantes a él. De allí la importancia de emplear tests que cuenten con normas estadísticas actualizadas y correspondientes a la región en la que el evaluado habita, obtenidas a partir de un grupo representativo de las personas típicas de esa edad y sexo (en el cap. 5 se ampliarán más estos conceptos).

Los *tests de criterio o de dominio*, en cambio, no emplean normas para comparar el desempeño o respuestas del individuo, sino que las valora según un criterio previamente establecido. Es decir que, por ejemplo, en una prueba elaborada para evaluar la presencia de síntomas depresivos, se determinará si las respuestas coinciden con el listado de síntomas de depresión que se ha tomado como base para comparar. Podría usarse, por ejemplo, el cluster de síntomas previsto en el DSM-IV (APA, 1995) o en la revisión de la CIE-10 (OMS, 2003), o el contemplado en cualquier otra descripción psicológica o psiquiátrica. De manera que nuestro sujeto obtendrá o no un diagnóstico positivo de Episodio Depresivo Mayor, o de Distimia, por caso, si sus respuestas coinciden con los signos y síntomas incluidos en el criterio establecido. Aquí la comparación ya no se efectúa sobre el promedio de los puntajes generados a partir de las respuestas de una muestra de sujetos y de su desviación típica (baremo) sino sobre el criterio o dominio especificado (síntomas depresivos según el DSM, la CIE o según cualquier otro criterio posible, ya sea descriptivo o teórico). Esto mismo puede aplicarse en un examen de admisión que compara las respuestas de los postulantes con ciertos criterios fijados de antemano, para así decidir su ingreso o no en un programa dado.

Finalmente, aunque no por ello como un asunto de menor importancia, debe tenerse presente que todo instrumento tiene su sentido en tanto y en cuanto esté ***enmarcado en un proceso de Evaluación Psicológica,*** cobrando valor e importancia en virtud del interjuego que sea posible establecer entre los resultados que arroje y el resto del material que se valorará (otros resultados derivados de otras técnicas psicométricas, proyectivas, entrevistas, informantes clave, etc.) sin perder de

vista el objetivo final de este proceso: la construcción de una descripción exhaustiva tendiente a generar una recomendación que llevará a tomar una decisión determinada. Todo esto se relaciona con el primer punto desarrollado en este capítulo: la *Evaluación Psicológica entendida como un proceso de toma de decisiones que funciona como una instancia consultiva, en tanto es solicitada por el interesado o por un tercero, para dar respuesta a alguna pregunta o consulta, o para indicar un curso de acción o solución a algún problema concreto.*

## 1.6 Ética del evaluador en Psicología: consideraciones básicas

Como último tema, haremos mención a las consideraciones éticas a ser tenidas en cuenta por el psicólogo evaluador. Comencemos por ciertas cuestiones que, a primera vista, aparecen como evidentes.

En líneas generales, diremos que el profesional que se desempeña en el área debe ser consciente de la influencia potencial que su trabajo puede tener en la vida de personas reales y concretas. Por ello, antes que nada, debe estar responsablemente formado y continuar su actividad en una constante inclinación al estudio, a la especialización y al aprendizaje. La plena conciencia de cuánto se sabe y de cuánto se ignora, de cuáles son las áreas de vacancia en sus conocimientos, sumada a una actitud responsable y honesta permitirán, por un lado, la capacitación permanente y, por otro, la derivación de aquellos casos para los que no se está habilitado, en virtud de limitaciones profesionales o personales.

Por otra parte, es fundamental tener en cuenta que la forma de conducirse en todo caso concreto es teniendo en claro cuáles son los derechos de todo examinado: tener acceso a la mejor atención que podamos brindarle –acompañada esta atención por las más adecuadas y mejores herramientas técnicas de que se disponga–, ser tratado con el mayor de los respetos, poder dar consentimiento informado para que la evaluación sea realizada, trabajar en un clima ameno, cordial y relajado, tener acceso a los resultados y recomendaciones surgidos de la evaluación en forma completa y adecuada a sus posibilidades intelectuales, emocionales y educativas, guardar los requisitos básicos de una estricta confidencialidad notificándole quiénes y de qué forma tendrán acceso a esos resultados (derivante, instituciones, familiares, etc.), contar con la autorización explícita de los padres para realizar evaluaciones a menores o a personas judicialmente insanas, con dificultades de comprensión o con patologías de alteración en el juicio de realidad, entre otros recaudos. Todos los cuidados anteriores implican una actitud activamente ética, así como de profundo respeto por el otro.

Asimismo, es importante no olvidar las diferencias culturales resultantes de la diversidad poblacional existente en nuestro país, de manera que aquellas no perjudiquen o beneficien espúreamente al sujeto en la valoración de su desempeño y no se vuelvan un obstáculo o un prejuicio para la correcta comprensión del caso; el profundo respeto por las peculiaridades culturales, religiosas, cosmovisionales, sexuales, sociales y de cualquier otra esfera del comportamiento humano, es la herramienta básica para la asunción de una postura abierta, humilde, empática, *contenedora* hacia el otro y carente de prejuicios –o, si tenemos en cuenta las afirmaciones de la Psicología Social sobre los prejuicios, al menos, con la menor cantidad de prejuicios posibles y siendo conscientes de ellos–. Estas cuestiones deben tenerse presentes a cada paso del proceso de la evaluación y, por lo tanto, también en el momento de elegir la batería de instrumentos a ser utilizada, a la hora de administrarla y en el



momento de puntuar los protocolos, al efectuar las interpretaciones y también al comunicarlas a los interesados. La actitud correcta consiste en una postura de total vigilancia respecto de estos tópicos, puesto que podemos estar trabajando frente a personas reales y concretas con especificidades macro o micro culturales que no hayamos detectado y, por lo tanto, tal vez estemos efectuando valoraciones que no corresponden a aquéllas, ya que estaremos mirando la realidad desde nuestra perspectiva cultural y no desde la que corresponde al sujeto.

Tomemos, por caso, la definición que Rohner (1984) hace del término *cultura*, entendida como un sistema compartido de símbolos y significados –complementarios y equivalentes– que se transmite de generación en generación y da origen a cierta estabilidad intergeneracional en una población determinada. La caracterización aquí esbozada hace más bien referencia a elementos ideales tales como valores, creencias, usos y costumbres, pero perfectamente puede hacerse extensiva a cualquier manifestación humana, ya sea aquella que involucre entes no tangibles –tales como los valores o las actitudes– así como las que impliquen productos reales y sensiblemente apreciables, como por ejemplo, bienes de producción, comportamientos observables, atuendos, ornamentos, entre las casi infinitas posibilidades a ser consideradas.

El tema cultural resulta de suma importancia, ya sea que hagamos un recorte más global o macrocultural, o una definición más microcultural. *Todo lo que el hombre hace es cultural* y pueden diferenciarse expresiones y grupos culturales si se piensa en distinciones fenoménicas entre nuestro país y otra nación, pero también pueden concebirse éstas si se comparan grupos de adolescentes de diversos estratos sociales dentro de un mismo barrio, adultos mayores con adultos jóvenes o la generación actual con la de sus tatarabuelos. Estas cuestiones serán abordadas en detalle en el cap. 5, pero valga esta sucinta introducción para vincular estos aspectos con el permanente cuidado que el evaluador debe mantener; atender a las peculiaridades culturales, respetarlas y tomarlas en cuenta implica una actitud ética por parte del examinador, tanto desde el punto de vista profesional cuanto personal.

Desde otro ángulo, es preciso prestar plena atención a las características particulares de cada persona, aquellas que la hacen única e irrepetible; para ello es importante mantener actualizados nuestros conocimientos sobre Psicometría, Técnicas Proyectivas, Psicología General, Psicología Evolutiva, Psiquiatría, Psicopatología, Psicología de la Inteligencia, de las Emociones, de la Motivación, de la Personalidad, relaciones de la Clínica Médica con la Psicología individual y Clínica, nociones básicas sobre Farmacología, entre otras áreas posibles. A la vez, el permanente acceso a los avances científicos y tecnológicos de actualidad –manuales de pruebas, textos dedicados a desarrollos científicos en Psicometría y Evaluación, catálogos de tests, artículos en revistas especializadas, asistencia a eventos científicos– se vuelve hoy en día, una actividad ineludible en la formación permanente a la que estamos aludiendo.

Por cierto, las mismas consideraciones que se han hecho hacia los sujetos evaluados deben contemplarse para con sus allegados o familiares, así como para con las personas que han solicitado la evaluación, para los informantes clave y para todo aquel que forme parte de este proceso. Este listado incluye, también, a las instituciones que se hallen implicadas, a sus actores y a sus representantes.

La comunicación de resultados, tanto la que se hace en el informe escrito cuanto la que se efectúa en la devolución oral, debe ser clara, directa, sin tecnicismos –o definiendo su significado con precisión y total detalle– , dosificada y adecuada a las posibilidades, cultura de origen, formación y características del destinatario, y debe

también basarse en un profundo respeto y consideración por éste. Jamás debe ser estigmatizadora, peyorativa o prejuiciosa y, desde ya, debe ser confidencial. Además, debe ser efectuada de manera responsable y hallarse inscripta en el proceso de evaluación psicológica en el que ha tenido lugar, teniendo sumo cuidado en comunicar los resultados en una forma que resulte contenedora y debidamente cuidada como para no producir ningún perjuicio en el sujeto, como por ejemplo en su autoestima, en sus expectativas de logro o en sus niveles de ansiedad. Toda devolución de resultados que pudiere introducir alguna modificación en estas áreas deberá ser acompañada de propuestas concretas –formales o informales– de intervención, dirigidas a brindar insumos prácticos, intelectuales, emocionales y sociales para afrontar estos resultados y estas indicaciones. Jamás una devolución de resultados debe generar un daño para el evaluado o para sus allegados.

Por otra parte, nunca debe olvidarse que el examinado es un *ser bio-psico-social*, que posee, por ende, las tres clases de componentes o atributos, sin perder de vista ninguno de ellos, pese a que nuestra mirada esté más entrenada para atender a los rasgos psíquicos antes que a los pertenecientes a las otras dos esferas. La postura adecuada y sensata es la que corresponde al profesional que no desecha ninguna de las tres dimensiones, reconociendo que no está capacitado para abordar todas y efectuando, de ser necesario, interconsultas con profesionales de otras áreas. Más bien la separación de tales *atributos* se basa en la división cuerpo-mente-sociedad que la instrucción universitaria nos ha transmitido tradicionalmente –así como los niveles educativos precedentes y aún el pensamiento vulgar de la vida cotidiana en nuestra cultura occidental– , y que deriva de un paradigma científico que supone esta escisión de manera tajante y artificial, en tanto que en la realidad el sujeto *es*, sin que sea posible escindir estos tres aspectos que se hallan íntimamente relacionados, formando parte de todos sus comportamientos en forma absolutamente integrada. Pensar en analizarlos separadamente ya implica una postura frente a la posibilidad de distinguirlos; tengamos en cuenta que se trata de distinciones sobre los fenómenos que no tienen que ver con la forma en que éstos se dan en la realidad, sino con nuestra manera de conceptualizarlos y de *aislarlos* para entenderlos y estudiarlos mejor.

No es el propósito de este apartado el brindar un exhaustivo listado de todos los puntos que hacen a la ética del psicólogo que se desempeña en esta área de trabajo, sino simplemente brindar algunos lineamientos generales que corresponden al profesional en general y a nuestra tarea en particular. Para ampliar estos tópicos se recomienda la lectura atenta de publicaciones vinculadas a la ética del evaluador (Asociación Argentina de Estudio e Investigación en Psicodiagnóstico –ADEIP-, 1999, 2000; American Educational Research Association, American Psychological Association & National Council on Measurement in Education –AERA, APA & NCME-, 2004; Anastasi & Urbina, 1998; Casullo, 1996; Cohen & Swedlick, 2001; Hogan, 2004; International Test Comisión –ITC-, 2006).

Como reseña, teniendo en cuenta las Pautas Internacionales para el Uso de Tests (ADEIP, 2000; ITC, 2006), el Código de Ética del Psicodiagnosticador (ADEIP, 1999) y los Estándares para Tests Educativos y Psicológicos (AERA, APA & NCME, 2004), se ofrecen las recomendaciones contenidas en los párrafos siguientes.

Los lineamientos de la ITC –Comisión Internacional de Tests– (2006) suponen una actuación acorde con estándares éticos y profesionales, en cuanto al uso de los técnicas, a la permanente actualización del evaluador sobre el debate científico que tenga lugar en el área de especialización, así como en referencia al hecho de estar seguro de que las personas con las que se trabaja o para quienes se trabaja mantienen también



dichos estándares; implican también comportarse con respeto y sensibilidad hacia los examinados, hacia las personas de su entorno y hacia las instituciones involucradas en el proceso de evaluación, además de presentar la tarea de los evaluadores en forma positiva y equilibrada cuando ésta tiene lugar en ambientes relacionados con los medios de comunicación. Asimismo, recomiendan evitar desempeñarse profesionalmente en situaciones en las que existan intereses específicos en cuanto a los resultados de la evaluación, o en las que ésta pueda producir daños en la relación del evaluador con los examinados o clientes que contratan el servicio.

Tal como expresábamos al principio de este apartado, el especialista deberá asegurarse de contar con la competencia necesaria para efectuar evaluaciones mediante tests con plena idoneidad; así, deberá asumir la responsabilidad por su formación, entrenamiento y experiencia, trabajando de acuerdo con principios científicos y según la experiencia comprobada, manteniendo elevados estándares de competencia, ética y técnica; estará obligado a conocer los límites de la propia competencia y no actuará por fuera de ellos, manteniéndose constantemente actualizado respecto de los cambios y avances en relación con el uso y construcción de los instrumentos que emplee, incluyendo las renovaciones de normas y modificaciones en la legislación que pudieran influir en la utilización de estas herramientas.

Deberá, también, responsabilizarse explícitamente por el uso que haga de las técnicas de evaluación, ofreciendo estos servicios sólo con pruebas de adecuada calidad y que hayan sido debidamente adaptadas a las características de los examinados, y sólo cuando se encuentre debidamente preparado para su correcta administración, evaluación e interpretación. Asumirá, entonces, la responsabilidad por la batería elegida y por las recomendaciones formuladas, brindando información clara y suficiente acerca de los principios éticos y los aspectos legales implicados a todos cuantos participan del proceso. Se asegurará de que el contrato entre evaluadores y evaluados sea claro y que haya sido correctamente comprendido, a la vez que se mantendrá atento a cualquier consecuencia no prevista del uso de los instrumentos que utilice. Se esforzará por evitar cualquier daño o perjuicio hacia las personas examinadas, no prolongará innecesariamente el proceso de evaluación y jamás interferirá con el trabajo de otro colega.

En cuanto a la debida reserva que debe guardarse con respecto al material de las pruebas, el especialista en el área evitará el préstamo o venta de instrumentos a personas que no posean un título de Psicólogo acreditado –esta regla debería regir también para editoriales y librerías–. Se asegurará de que los materiales de examen, protocolos, cálculos de puntuaciones e informes se conserven en lugares seguros, que deberán estar a salvo de la intromisión de extraños, controlando siempre el acceso de toda persona a los mismos.

Se respetarán en todos los casos los derechos de autor, así como los acuerdos que rijan sobre cada instrumento, protegiéndose la integridad de los materiales, de manera que no puedan influir en el desempeño de los evaluados; así, se evitará la exposición pública de aquellos, de tal manera que su utilidad no quede alterada.

Tal como se indicaba en párrafos anteriores, se insistirá fuertemente en la estricta confidencialidad de los resultados, especificándose previamente a las partes implicadas quiénes tendrán acceso a los mismos, y limitándose dicha información sólo a aquellos que tengan la estricta necesidad de conocerla. Deberán, asimismo, obtenerse autorizaciones explícitas antes de darlos a conocer a otras personas que posteriormente se juzgue deban acceder a aquéllos; se protegerán los datos archivados de manera que solamente sean accesibles a quienes se arroguen justamente tal derecho,

estableciéndose pautas claras acerca del tiempo que se esperará para eliminarlos o destruirlos, de acuerdo con los Códigos de Ética y legislación vigentes en cada país o provincia.

Se suprimirán datos identificatorios -sin que medie pedido de los interesados- en el caso de bases de datos destinadas a la construcción de baremos o que se confeccionen para servir a investigaciones. La exposición de casos en congresos o seminarios deberá eliminar cualquier posibilidad de identificar a los evaluados o sus allegados.

Siguiendo la idea esbozada párrafos atrás, para que el uso de los instrumentos resulte adecuado, deberán ser empleados por usuarios competentes, capaces de ofrecer una justificación razonada para la elección de los mismos, habiendo efectuado un riguroso y exhaustivo análisis de las necesidades de las partes consultantes, las condiciones o trabajo para los que se usará la evaluación y, de ser necesario, de las categorías diagnósticas implicadas en los resultados. El evaluador idóneo deberá haber comprobado que los tests elegidos miden, efectivamente, atributos o rasgos que sean correlatos de comportamientos relevantes en el contexto en el que se llevarán a cabo las inferencias efectuadas; con este mismo fin, deberá proveerse de fuentes de información adicionales, así como de investigaciones sobre los instrumentos que usará, sopesando ventajas e inconvenientes involucrados en el uso de estas herramientas por sobre otras fuentes.

En esta línea de acción, elegirá escalas adecuadas desde el punto de vista técnico y según cada situación particular; para ello, deberá examinar concienzudamente la totalidad de la evidencia disponible antes de decidir la batería de instrumentos a ser administrada, verificando que esa documentación contenga información suficiente acerca de los grupos normativos empleados, sobre la dificultad y características del contenido de los ítems, acerca de estudios empíricos de fiabilidad, evidencias de validez y potencial aplicabilidad de los resultados a situaciones concretas, así como ausencia de sesgos culturales para los grupos en los que se usará, además de otros aspectos prácticos varios a ser considerados.

En particular referencia a la validez (ver cap. 2), el especialista responsable usará sólo escalas que cuenten con adecuadas evidencias empíricas, demostradas mediante una documentación completa y bien fundamentada. Jamás debe aceptarse el uso de un instrumento sobre la base de su *validez aparente o de facies* (ver cap.2), o en virtud de las recomendaciones de otros usuarios o de quienes tengan intereses comerciales implicados.

Por otra parte, el examinador deberá responder en forma clara y precisa a las preguntas de las personas que participan en el proceso de evaluación -examinados, familiares, jueces, gerentes, representantes legales, educadores, entre otros actores posibles-, brindando información suficiente para que les sea posible comprender los fundamentos de la elección de la batería.

Volviendo a los aspectos culturales que se abordaban algunos párrafos atrás, deberá prestarse especial atención a las cuestiones relacionadas con el sesgo cultural -funcionamiento diferencial de los ítems en distintos subgrupos, que los tests no sean imparciales en todos los grupos y subgrupos en los que puedan utilizarse, que los constructos evaluados sean relevantes para cada uno de los grupos examinados, que las diferencias de rendimiento en los distintos grupos sean explicadas por diferencias reales en las variables evaluadas y no por razones de pertenencia a sectores minoritarios o con determinadas desventajas, es decir, que se minimicen las diferencias grupales no relacionadas con el objetivo principal de la evaluación y, finalmente, que se aporten datos sobre la validez de los resultados arrojados por la prueba en distintos



subgrupos-. Estos lineamientos se enumeran aquí a los fines de introducir su importancia ética, pero serán desarrollados en detalle en los caps. 2 (sesgo) y 5 (adaptación de tests a distintas culturas).

El usuario responsable deberá, asimismo, prestar especial atención a la sensibilidad que los autores hayan tenido sobre cuestiones idiomáticas y lingüísticas en general: los aspectos de contenido deberán haber sido muy cuidados en las adaptaciones transculturales de estos instrumentos, a la vez que no deberán ser meras traducciones lineales, sino que deben haber sido elaboradas con la metodología más rigurosa y adecuada (ver cap.5).

Continuando con el tópico idiomático, también deberá preverse que los aplicadores sean perfectamente capaces de comunicarse en el idioma en que se aplica la prueba; a la vez, los evaluados deberán ser examinados con la versión más adecuada a la lengua con que se manejan más cómodamente. Todo recaudo especial que pueda tomarse para el empleo de instrumentos en personas con capacidades especiales debe ser fuertemente recomendado, meditado y puesto en práctica con el mayor de los esmeros.

Huelga decir que las sesiones de administración deben ser cuidadosa y minuciosamente preparadas, de acuerdo con los requisitos de una buena praxis y con el establecimiento de un adecuado *rapport*, haciendo eje en el especial cuidado de los derechos del examinado, y en el curso de una relación interpersonal y no por medio del teléfono, el correo postal o electrónico o, por caso, Internet.

Permanentemente debe intentarse reducir la ansiedad del evaluado y evitar la generación de ansiedad innecesaria, eliminar potenciales fuentes de distracción, asegurarse de que los materiales se encuentren en buen estado, prever con cuidado las dificultades que puedan surgir, apegarse estrictamente a las consignas y hacer ajustes en caso de alguna discapacidad o inconveniente puntual; las instrucciones deben impartirse en forma clara y pausada, respetándose los tiempos de administración adecuados, y asegurándose de que todos los examinados reciban la supervisión y asistencia que sean necesarias; para ello, ha de entrenarse adecuadamente a los ayudantes, asegurándose de que la administración no quede desatendida o de que las personas evaluadas no queden sujetas a distracción o sin asistencia en los casos en que experimenten desazón o ansiedad excesiva ante la evaluación. Al finalizar la tarea de administración, el responsable de la misma deberá verificar que los materiales estén completos y sin datos faltantes. Desde ya que las condiciones ambientales de luminosidad, temperatura, ruidos, confort, elementos de trabajo y tranquilidad también deberán ser objeto de preocupación.

En cuanto a la puntuación y análisis de los protocolos, estas tareas se realizarán con precisión y cuidado, respetando las indicaciones de los manuales -especialmente cuando puedan entrar en juego juicios de valor psicopatológico-, efectuando minuciosamente las transformaciones de puntuaciones que permitan la adecuada lectura de los mismos y eligiendo el tipo de escala más conveniente a las características del test y de las variables evaluadas. Se evitará por todos los medios arribar a conclusiones erróneas a causa del empleo de baremos desfasados o inadecuados para los evaluados; se examinarán y revisarán los resultados para detectar equivocaciones o anomalías, debiéndose, además, describir e identificar las normas, fórmulas y conversiones realizadas, integrando métodos de análisis cualitativos y cuantitativos.

La interpretación de los resultados se asentará sobre sólidas bases teóricas y técnicas, y será acorde con las características y circunstancias del sujeto evaluado, avalándose con la documentación adecuada y actualizada sobre la prueba, sumándose a

todo lo anterior un acabado conocimiento de las escalas, las puntuaciones y demás elementos técnicos incluidos. En el informe y la devolución oral de resultados se cuidarán las generalizaciones que excedan los límites del rasgo evaluado, la teoría de base y/o los propósitos del instrumento, teniendo en cuenta la confiabilidad, el error de medida y las evidencias de validez de los resultados a la luz de las características de los sujetos examinados (para comprender mejor estos conceptos, se recomienda ver caps. 2, 3 y 5).

El especialista formado, además, deberá concientizarse sobre los prejuicios y estereotipos negativos o positivos que pueden afectar sus interpretaciones – ya sea para sobrevalorar o para subvalorar el desempeño del sujeto–, dados por variables tales como sexo, edad, cultura de origen, religión, preferencias sexuales o condición social, entre otras.

Tampoco deben perderse de vista las variaciones introducidas en los procedimientos estandarizados, así como las limitaciones del evaluador en cuanto a su experiencia previa con cada instrumento en particular; de la misma manera, debe sopesarse siempre la experiencia del evaluado con este tipo de tests, con el fin de realizar una ajustada valoración sobre sus respuestas en la situación presente.

Si bien es cierto que el psicólogo puede recurrir a especialistas de otras disciplinas para obtener los cómputos y puntuaciones, la evaluación como totalidad, de principio a fin, sigue siendo de su total y absoluta responsabilidad, así como la interpretación y comunicación de los resultados. Siguiendo esta última idea, dicha comunicación se hará en forma clara y precisa a todos los involucrados que deban tener acceso a esos resultados, dejando de lado prejuicios o estereotipos y con el consentimiento de las partes implicadas, tal como se comentó antes; el nivel de lenguaje utilizado deberá ser adecuado a los lectores/receptores, con el objeto de minimizar la posibilidad de interpretaciones erróneas.

Se incluirán las debidas fundamentaciones de los diagnósticos vertidos y/o de los resultados obtenidos, incluyendo un resumen claro de los mismos, así como recomendaciones de cursos de acción concretos toda vez que ello corresponda. Los evaluados y sus allegados deben recibir la información en forma positiva y constructiva, evitando infringir cualquier daño, por mínimo que éste fuera. Nunca deberá suministrarse información sobre las respuestas esperadas puesto que ello podría invalidar procedimientos de evaluación ulteriores.

El especialista deberá siempre estar muy atento a la información verbal y comportamental que proporcionen el evaluado u otros referentes, con la finalidad de que no se produzca un uso inadecuado de la misma. Los datos puntuales proporcionados por el evaluado pertenecen a la esfera del secreto profesional y solamente se informará a los derivantes o familiares acerca de diagnósticos, problemáticas generales o recomendaciones, sin aludir directamente a los contenidos concretos que el examinado confió al examinador.

Por su parte, el *Código de Ética del Psicodiagnosticador*, elaborado en nuestro país (ADEIP, 1999), con un espíritu muy similar al exhibido por las normas de la International Test Commission (ITC, 1996), proponen la regulación ética de la tarea concreta del psicólogo que se desempeña en el ámbito de evaluación, haciendo eje en doce vectores: 1) que la evaluación, el diagnóstico y las intervenciones del área deber ser efectuadas en un contexto estrictamente profesional, 2) la necesaria competencia del evaluador especializado y el uso apropiado que debe hacerse de cada evaluación e intervención, 3) las cláusulas referidas al secreto profesional, 4) las cuestiones relativas a la construcción de tests, 5) el buen uso que debe hacerse de la evaluación en general y con poblaciones especiales, 6) la adecuada interpretación de los resultados de la evaluación, 7) el deber de no promover la aplicación de técnicas de evaluación por parte de personas no calificadas, no habilitadas por un título universitario idóneo y sin una preparación especial ulterior a su formación de grado, 8) la vigencia de los tests, sus baremos y revisiones, 9) la necesaria comprobación de la adecuación e idoneidad de los servicios informatizados de puntuación e interpretación que son, en última instancia, responsabilidad del evaluador que los contrate, 10) el mantenimiento de la seguridad de los tests, de sus materiales, protocolos, datos para investigación y reserva en cuanto a las respuestas correctas o esperadas, especialmente con sujetos que van a ser evaluados, 11) cuestiones ya mencionadas en cuanto a la comunicación de los resultados de la evaluación, y 12) restricciones en cuanto a la difusión en medios de comunicación, ya que ello puede afectar las aplicaciones futuras de los instrumentos. No abundaremos aquí en el detalle de los doce puntos anteriores puesto que ya han sido abordados en forma ampliada en los párrafos precedentes.



Si bien no nos extenderemos aquí en los recaudos que deben tomar los constructores de tests o los investigadores que realizan adaptaciones de los mismos a poblaciones distintas de la original, los preceptos básicos que se aplican a esa esfera son, en todo caso, trasladables al usuario. Los estándares internacionales (ADEIP, 2000; AERA, APA, NCME, 2004; ITC, 2006) hacen hincapié en que los autores y editores proporcionen al usuario **todas** las especificaciones existentes en cuanto a evidencias de validez, estudios de confiabilidad, construcción de baremos, sesgos, adaptaciones culturales, estandarización, entre otros, para hacer un uso responsable, ético y teórica y técnicamente correcto de los instrumentos de evaluación. Dado que este texto no está dirigido a diseñadores sino a usuarios, la idea general que aquí se quiere transmitir se vincula a que cuanto más capacitado se encuentre el evaluador especializado, más exigente se volverá con las técnicas, con sus manuales y con las especificaciones en ellos contenidas y esto redundará en beneficio de los examinados, así como de todos los personajes e instancias implicados en el proceso de evaluación. Cuanto más formado se halle un especialista y cuanto más vigilante se mantenga con respecto a la calidad de los tests, así como hacia las normas éticas a ser cumplidas en todos los casos, mejor desarrollará su labor y mayor será el beneficio reportado a todos los interesados.

De esta manera, los manuales deberán consignar los criterios y pasos seguidos en la construcción de los instrumentos, brindando la información adecuada para su utilización, evidencias de su validez –y, por ende, usos potenciales, aplicaciones, sujetos sobre los que se determinaron estos resultados, resultados concretos y métodos empleados–; asimismo, deberá desecharse el empleo de cualquier prueba cuyo manual no informe sobre las características, utilidad y limitaciones de esa escala, su teoría de base y la operacionalización del constructo a ser evaluado, la descripción de la o las muestras empleadas en la tipificación o estandarización, la selección que se hizo del contenido de los ítems, de su formato y evidencias sobre su confiabilidad y error de medida. El usuario debe notar a cada paso de su lectura del manual y de artículos científicos relacionados, que los autores originales han cuidado minuciosamente la calidad científica y física del material, así como la información contenida en aquel. Esta impresión se verá reforzada en tanto y en cuanto el circuito comercial que se mueve alrededor de un instrumento dado no desvirtúe los principios establecidos en estas pautas. El examinador deberá poder hallar, también, en estos materiales, información sobre escalas, subescalas y comparabilidad de las mismas, junto con las especificaciones completas sobre la administración, puntuación e interpretación de

los resultados. Deberá adjuntarse, además, documentación complementaria que avale el test desde los puntos de vista metodológico, teórico y técnico.

Refiriéndonos ahora a las revisiones, adaptaciones y actualizaciones que se hagan de las pruebas, los estándares internacionales intentan hacer comprender que los mismos deben sufrir revisiones periódicas, permanentes ajustes técnicos de forma, actualización de contenidos o forma de aplicación, de baremos, de estudios de confiabilidad y de validez, entre otras. Un instrumento que se ha mantenido inalterado y sin revisiones a lo largo de los años, seguramente, no está tomando en cuenta las variaciones que las variables psicológicas experimentan a lo largo del tiempo histórico y de la disparidad cultural y/o geográfica. Se deberá encontrar documentación sobre la equidad y el sesgo cultural propio de la escala en diversos grupos en los que haya sido probada, así como las peculiaridades relacionadas con el examen de personas con bagajes lingüísticos diversos, y la evaluación de individuos con capacidades especiales.

Por último, deberá disponerse, también, de un exhaustivo desarrollo sobre los derechos y reponsabilidades de los usuarios, y de cuestiones formales, técnicas, metodológicas y teóricas vinculadas a la evaluación en los ámbitos clínico, forense, educativo, laboral, de evaluación de programas y de políticas públicas.

Todos estos recaudos guardan la intención de proteger a los evaluados, a sus allegados, a todas las personas e instituciones involucradas en la situación de evaluación y, por qué no, a la práctica concreta de la Evaluación Psicológica en sí. Cuanto más formado y competente sea el especialista del área, más exigente será para con los materiales de los tests y para con su propio quehacer que, definitivamente, impactan en la vida de las personas de una forma o de otra. En última instancia, la idoneidad profesional es, también, una cuestión de ética.

*Capítulo 2*

# *La validez y los instrumentos psicométricos*

Mercedes Fernández Liporace

**Contenidos temáticos**

- **Validez (concepto)**
- **Aspectos de la validez relacionados con el contenido**
- **Aspectos empíricos de la validez (validez de criterio)**
- **Aspectos de la validez vinculados al modelo teórico**
- **Aspectos de la validez vinculados con las características formales de la prueba**
- **Sesgo y error sistemático**

## 2.1 El concepto de validez

Hasta hace algunos años, hubiéramos comenzado este capítulo afirmando que existen, básicamente, tres elementos fundamentales que permiten juzgar la calidad de una técnica psicométrica: su *confiabilidad*, su *validez* (Cohen & Swerdlik, 2001) y su *capacidad discriminativa*, si bien otros autores enumeran condiciones que resultan altamente recomendables, relacionadas con las instrucciones, los ítems y otros aspectos, especialmente los formales (Martínez Arias, 1995). Los clásicos manuales de psicometría mencionan esos tres atributos como los mínimos necesarios a ser examinados con el fin de valorar si un instrumento dado nos aporta la información que necesitamos conocer, con el agregado de que, en base a ellos, sabremos si el test ha sido construido y estudiado según una metodología científica seria. Vayamos por partes.

Tal como comentábamos en el capítulo anterior, cuando definíamos y caracterizábamos los tests, al hablar de ***discriminación***, nos referíamos a la **capacidad de un instrumento (en realidad, a la capacidad de sus ítems) para captar diferencias individuales en la variable que está siendo medida.** Decíamos entonces que éste es el sentido último de las escalas psicométricas: la determinación de diferencias en un atributo dado entre distintas personas (Martínez Arias, 1995). Si los ítems de una prueba poseen una baja capacidad discriminativa, entonces, su utilidad se reduce considerablemente. Asimismo, los autores tienen la obligación de proporcionar a los usuarios resultados empíricos - determinados a partir del rendimiento de sujetos concretos - sobre tal poder discriminativo.

En segundo término, la ***confiabilidad* de una prueba se refiere a la *confianza que podemos tener en los resultados* que arroja**, confianza que puede ser examinada desde varios aspectos o aristas, tal como veremos en el capítulo 4, pero que principalmente se dirige a valorar cuánto error existe en la medición, asumiendo, desde ya, que siempre se incluyen errores en la misma, ya que no es posible concebir medición alguna que esté completamente libre de error (Martínez Arias, 1995); (ver cap. 4).

En tercer lugar, debemos ocuparnos de la *validez* de la escala, tema de este capítulo. Tradicionalmente, **se alude a la *validez* mediante la pregunta referida a *qué mide la técnica*** (Anastasi & Urbina, 1998) **y cómo lo mide** (Casullo, Figueroa & Aszkenazi, 1991). Que un test haya sido nombrado como prueba de inteligencia, o de memoria, o de personalidad, no implica necesariamente que mida la o las variables que figuran en su denominación. El autor debe proporcionar pruebas empíricas –resultados verificados, verificables y replicables por otros investigadores– que demuestren que, efectivamente, se está midiendo la inteligencia y no la memoria, o la persistencia o cualquier otra variable no identificada, diferente de la que se pretende evaluar en el propósito explícito declarado en el manual del instrumento.

En Psicología, este problema de medición se vuelve particularmente importante puesto que la inmensa mayoría de las veces –si no todas– un comportamiento puede ser explicado por la interacción de varios factores que lo determinan, que hemos dado en llamar variables, conceptos o constructos, según nos ubiquemos en un punto de vista más empírico o más teórico. Por ejemplo, el rendimiento escolar de un alumno estará dado por su inteligencia, pero también se verá propiciado por el interjuego de muchos otros factores, tales como su motivación, sus habilidades en cuanto a las destrezas más requeridas en el contexto escolar, su ansiedad, la relación establecida con el docente y con sus pares, entre muchas otras variables a ser tenidas en cuenta. El mismo razonamiento puede aplicarse a otros comportamientos o desempeños particulares. Asimismo, el tipo y contenido de los ítems que se incluyan en el instrumento dependerán de la definición específica que el enfoque teórico bajo el que se trabaja para elaborar la técnica hace de la variable a medirse (ver cap. 1). Por estas razones, **jamás debe darse por sentado que la denominación de la técnica responde exactamente al constructo que se pretende evaluar. Debemos contar con pruebas empíricas que lo hayan verificado, aportadas por los autores del test mismo, así como por otros investigadores que lo hayan analizado ulteriormente.** Estos resultados empíricos, que serán el producto final de una investigación planificada y desarrollada con todo rigor metodológico, serán uno de los fundamentos básicos que contribuirán a asegurar la calidad de la técnica que nos ocupe, puesto que se refieren a uno de los atributos que certifican esta calidad: la validez de las puntuaciones que la prueba permite obtener.

Así, se comprende que **la validez de un instrumento se refiere a que esa herramienta sirva para medir aquello que intenta medir** (García Cueto, 1993). Aunque existe una multiplicidad de definiciones sobre este concepto, fácilmente es posible apreciar que todas ellas se relacionan con la idea de que el test debe medir, efectivamente, el constructo que se ha propuesto o que ha prometido. Algunas de ellas son:

- El grado en el que se predice o mide un criterio dado (Lord & Novick, 1968).
- La verificación de que el instrumento sirve para aquello que se pretende de él (Yela, 1987).
- El soporte adecuado a las inferencias que se formulen en base a las puntuaciones obtenidas a partir de la prueba, establecido mediante un proceso de investigación cuidadosamente desarrollado (García Cueto, 1993).



- La posibilidad de utilizar y aprovechar las inferencias realizadas a partir de las puntuaciones arrojadas por esa escala, dentro de ciertas condiciones previamente establecidas, con un objetivo dado (Crocker & Algina, 1986).

Volviendo ahora a los primeros renglones de este capítulo, afirmábamos que hasta hace algunos años, hubiéramos aseverado que los tres elementos que permiten juzgar la calidad de una técnica psicométrica son su *capacidad discriminativa* (Hogan, 2004), su *confiabilidad* y su *validez* (Cohen & Swerdlik, 2001). Introduzcamos ahora una pequeña variación en esta oración antes de comenzar a desarrollar el tema *validez*.

Desde hace un tiempo, y a partir de las últimas ediciones de los Estándares para la Evaluación Educativa y Psicológica (AERA, APA & NCME, 2004) se ha producido cierto viraje en la terminología empleada en Psicometría, que no es meramente un cambio de palabras, sino que refleja una manera algo diferente de comprender o de definir los conceptos antes abordados. *Ya no hablamos de discriminación, confiabilidad y validez como atributos inherentes al test, ya que ello crea cierta sensación de invariabilidad, de permanencia o de fijeza. Actualmente nos referimos a la* ***confiabilidad y validez de los resultados arrojados por el test****, así como a la capacidad discriminativa de sus ítems* ***en tales o cuales sujetos, que poseen tales y cuales características determinadas*** (AERA, APA & NCME, 2004). Ello significa varias cosas; veamos las más importantes. En primer término, y como ya anticipábamos, que la validez y la confiabilidad de los resultados no vienen dadas con el instrumento, sino que se hallan sujetas a ciertas condiciones, referidas especialmente a la *variabilidad* de la muestra de sujetos que se ha utilizado para poner a prueba, justamente, la validez y confiabilidad de los resultados, así como la discriminación de los ítems que componen la escala.

En segundo lugar, también quiere decir –como antes resaltábamos– que tales atributos vinculados a los resultados arrojados por el instrumento deben ser empíricamente determinados mediante investigaciones científicas meticulosamente planificadas y desarrolladas, a la vez que tales estudios deben ser replicados y/o rediseñados a intervalos temporales relativamente cortos, así como cada vez que la prueba se traslade de un contexto cultural a otro.

Detengámonos un instante para desmenuzar los dos párrafos anteriores. Decíamos que **validez, confiabilidad y capacidad discriminativa no son atributos que vienen dados con el test o directamente derivados de él, sino que están sujetos a determinadas condiciones**. Cuando un instrumento psicométrico se estandariza (ver cap. 1), se trabaja este proceso de tipificación con muestras de sujetos que reúnan características homogéneas a las personas a las que se ha destinado esa prueba. Es decir que si se trata de una escala para niños, la estandarización o tipificación se realizará sobre una muestra infantil, que debe ser representativa de la población de niños a los que potencialmente se piensa administrar luego ese instrumento, en el ámbito de aplicación (ver cap.1). Análogamente, si la escala se destina a adolescentes escolarizados de clase media residentes en la ciudad de Buenos Aires, la muestra de sujetos sobre la que se realicen las investigaciones sobre la validez y confiabilidad de los resultados debe estar compuesta, efectivamente, por adolescentes escolarizados de clase media residentes en Buenos Aires, que sean representativos de la mayoría de los adolescentes *típicos* que reúnan estas características, especialmente teniendo en cuenta que sean *típicos* en relación a la variable evaluada por el test. Así, si la escala mencionada se ocupa de medir la inteligencia académica de los sujetos, la muestra no debería estar compuesta por una mayoría de adolescentes que registren un desempeño escolar sobresaliente o, a la inversa, un rendimiento muy bajo, sino que debería estar

formada por alumnos adolescentes de Buenos Aires que tengan rendimientos académicos variados, y que los mismos se distribuyan de la misma manera que en la población. De otra forma, se estarían tomando en cuenta resultados derivados de desempeños pertenecientes a personas muy especiales, de baja o escasa aparición en la población a la que destinamos la prueba, y ello empañaría la correcta interpretación de aquéllos. Por supuesto que es posible efectuar estos estudios de calidad de los tests en poblaciones especiales, pero deben tomarse como tales, y no como derivadas de investigaciones desarrolladas con personas típicas de determinada población o universo.

Una vez comprendido este punto, es importante tomar en cuenta, entonces, que las características distintivas de cada muestra –o población, si efectivamente la muestra es, definitivamente, representativa de la población a la que pertenece– afectarán el desempeño o las respuestas de los sujetos que la componen. Si deseamos estudiar la validez de los resultados obtenidos mediante un test de personalidad en sujetos pertenecientes a la población general y también en pacientes psicóticos, deberíamos encarar dos estudios separados: uno con una muestra representativa de personas provenientes de la población general y otro con pacientes diagnosticados como padeciendo determinados trastornos psicóticos, ya que tomar ambos grupos –tan diferentes entre sí– como uno solo y homogéneo podría producir un serio error en la interpretación y valoración de los resultados. De la misma manera, si la idea es desarrollar investigaciones sobre la calidad de un test de inteligencia en individuos con determinados talentos especiales, en la muestra seleccionada deberían incluirse personas típicamente representativas de quienes integran la población con ese talento dado, debiendo planificarse estudios independientes si se trata de examinar la calidad de la escala en personas con inteligencia promedio o en individuos con desempeños intelectuales muy disminuidos. Es fácil comprender estos conceptos si podemos imaginarnos que el nivel intelectual de un sujeto afecta su rendimiento en tareas vinculadas con esa variable, así como también el estatus de paciente o no paciente de otra afectará sus respuestas en un test de personalidad.

Ahora bien, más allá de estos ejemplos, en Estadística y Psicometría se habla de una característica específica de la distribución de puntuaciones obtenidas por la muestra de sujetos en un test, que es la *variabilidad* o *dispersión* de las mismas. Recordemos que, en líneas generales, las medidas de variabilidad nos indican si las puntuaciones o valores obtenidos por los distintos sujetos que componen la muestra están próximas entre sí o si por el contrario se encuentran muy dispersas (Botella, León & San Martín, 1997). Dicho de otro modo, si la variabilidad es pequeña, este dato nos informa que la mayoría de las puntuaciones se ubican cerca de la media o promedio (suma de puntuaciones obtenidas por todos los sujetos de la muestra, dividida por el número de casos o sujetos tenidos en cuenta). Por el contrario, si la dispersión es grande, podemos inferir que las puntuaciones se encuentran, precisamente, dispersas en relación a la media. Una media de 5 puntos en los exámenes finales, obtenida por dos muestras de alumnos diferentes, la primera con una desviación típica de 0.2 y la segunda con un desvío de 3.1, nos indicará que en ambos grupos el promedio de calificaciones obtenidas es el mismo, pero en el primero, un desvío pequeño informa que las puntuaciones se hallan muy juntas, que son bastante semejantes entre sí y que se reúnen cerca de la media; estos estudiantes se parecen entre sí en cuanto a sus desempeños. El otro grupo, con una desviación de 3.1, en cambio, exhibe rendimientos menos semejantes entre los individuos, es decir, más dispersos o heterogéneos: la distribución de calificaciones tendrá una mayor variabilidad, incluyendo



notas más altas, más bajas y medias, en tanto que en el primer grupo tal variabilidad será mucho menor, pareciéndose las calificaciones entre sí.

Justamente es esta variabilidad la que afecta los resultados que se obtienen en los estudios desarrollados para determinar la validez y la confiabilidad de las puntuaciones arrojadas por un test (Martínez Arias, 1995). A este tema nos referíamos, en concreto, cuando decíamos que la validez de las puntuaciones se encuentra influida por la variabilidad de la muestra sobre la que esa validez se determina: no registrarán iguales resultados las muestras de sujetos que presenten mayor dispersión en las puntuaciones, si se las compara con muestras cuyos rendimientos hayan sido más similares entre sí.

Apelaremos ahora a un ejemplo muy sencillo que ayudará a comprender este concepto. Se dispone de resultados obtenidos en un test por dos muestras de sujetos, que daremos en llamar A y B, compuestas cada una por 5 casos. Ellos han obtenido las siguientes calificaciones:

| Muestra A | | | | |
|---|---|---|---|---|
| *Sujeto 1* | *Sujeto 2* | *Sujeto 3* | *Sujeto 4* | *Sujeto 5* |
| 7 | 7 | 7 | 7 | 7 |

| Media muestra A | Desvío estándar muestra A |
|---|---|
| 7,00 | 0 |

| Muestra B | | | | |
|---|---|---|---|---|
| *Sujeto 1* | *Sujeto 2* | *Sujeto 3* | *Sujeto 4* | *Sujeto 5* |
| 10 | 9 | 8 | 3 | 5 |

| Media muestra B | Desvío estándar muestra B |
|---|---|
| 7.00 | 2.91 |

Como puede apreciarse, ambas muestras han obtenido una media de calificaciones de 7 puntos, pero la muestra A registra una desviación típica igual a 0 y la B, una de 2.91. Los valores mínimo y máximo observados en la muestra A coinciden en 7 puntos, ya que las calificaciones han sido todas iguales a 7, en tanto que en la muestra B se aprecia un mínimo de 3 y un máximo de 10. Así, puede advertirse que la dispersión de notas presentadas por ambas muestras difiere a pesar de que las medias sean iguales. Es sencillo, entonces, imaginarse las diferencias de rendimiento si pensamos en una muestra en la que todos los alumnos obtienen la misma calificación y en otra, en la que aparecen valores de 10, 9, 8, 5 y 3. A estas distinciones nos estamos refiriendo cuando aludimos al concepto de *variabilidad muestral*. Y puesto que los estudios de validez se efectúan sobre muestras compuestas por personas reales, con respuestas y rendimientos concretos y reales, es que tales resultados de dichas investigaciones se encontrarán directamente afectados por esta característica de la muestra de sujetos, la cual no es ni más ni menos que el resultado del conjunto de atributos psíquicos específicos y distintivos de un grupo de individuos dado: estas diferencias, que parecen distinciones numéricas *extrañas*, no son ni más ni menos que diferencias registradas en el rendimiento real observado por personas reales y que, por lo tanto, afectarán los estudios que se hagan sobre tales sujetos.

Para nosotros, psicólogos y futuros psicólogos, acostumbrados a comprender intuitivamente que distintos atributos psíquicos determinan puntuaciones disímiles en un test, el ejemplo anterior puede resultar de utilidad para apreciar *numéricamente* tales diferencias.

Recapitulando, habíamos mencionado que las propiedades psicométricas vinculadas a los resultados arrojados por el instrumento –***evidencias de validez*** en este caso– deben ser empíricamente determinadas mediante investigaciones científicas meticulosamente planificadas y desarrolladas, a la vez que dichas investigaciones deben ser replicadas y/o replanteadas a intervalos temporales relativamente cortos, así como cada vez que la prueba se traslade de un contexto cultural a otro. Justamente en virtud de lo que referíamos en los párrafos precedentes –las características distintivas de cada grupo de sujetos, es decir, la dispersión de sus rendimientos, influirán en los resultados de los estudios acerca de la validez de las puntuaciones arrojadas por los tests en esas muestras de sujetos–, es que los estudios de validez deberán repetirse toda vez que se varíe la población –y por lo tanto la muestra– sobre la que se ha trabajado; en toda oportunidad en la que se traslade el uso potencial del instrumento a una nueva población humana deberán renovarse los ***estudios que aporten evidencias de validez*** (Martínez Arias, 1995).

Por último, y antes de comenzar a tratar más a fondo el tema validez, repetiremos que, de acuerdo con los estándares de construcción de pruebas vigentes en la actualidad (AERA, APA & NCME, 2004), haremos notar otro cambio de terminología que implica, a la vez, un viraje en el enfoque que antes se tenía sobre el concepto de validez, cambio que se relaciona con las expresiones destacadas en negritas cursivas en el párrafo anterior. Antes de las últimas ediciones de estos estándares –tal como puede observarse en todos los manuales y textos de psicometría con algunos años de antigüedad– se hablaba de *la validez de los instrumentos psicométricos*. Ya hemos dejado aclarado el punto de que *actualmente nos referimos a la validez de las puntuaciones arrojadas por una prueba*, pero además se especifica que **ya no se trata de la validez del test, sino de *evidencias de validez* de tal o cual tipo, aportadas por tal o cual investigación desarrollada sobre tales o cuales sujetos.**

Esta diferenciación, que también suena como un simple cambio de palabras, implica varias distinciones conceptuales mayores: como ya se comentó antes, la validez no es ya considerada un atributo estático de la escala, sino que es pensada como un resultado que corresponde a una investigación científica efectuada en un momento concreto y con personas reales y concretas, por lo que sus resultados son provisorios, temporarios, y sujetos a refutación y verificación constantes. Y en segundo término, se tiene muy en cuenta –como veremos en los próximos apartados– que al diferenciarse varios tipos de **validez de las puntuaciones obtenidas** por medio de una técnica, no se puede concluir si una escala es válida o no, sino que debe afirmarse que *se han obtenido evidencias de validez de tal o cual tipo en las puntuaciones derivadas de tal o cual instrumento* –ya se verán en las páginas siguientes– *en tales sujetos y bajo determinadas condiciones.*

Así, a la luz de las aclaraciones precedentes y volviendo a la definición inicial de validez –qué variable mide un instrumento y en qué grado (Hogan, 2004)–, vale la pena preguntarse no ya si un test presenta validez –en abstracto–, sino si ha obtenido evidencias de validez predictiva (por ejemplo, ya que, como anticipamos, existen varios aspectos a considerar en este concepto) en sujetos adolescentes no escolarizados, residentes en La Quiaca, utilizándoselo con fines clínicos. Los apartados siguientes servirán para esclarecer aún más estos conceptos



Distintos tipos de validez

Una vez introducidos al concepto de validez y habiendo efectuado las aclaraciones precedentes, pasemos a examinar las distintas aristas involucradas en el tema.

Existen diversas nomenclaturas sobre los aspectos a ser tenidos en cuenta en relación con el concepto de validez o, como expresamos hoy en día, con las diferentes *evidencias de validez que pueden ser aportadas sobre los resultados arrojados por un instrumento psicométrico* dado. No referiremos a las más comunes y difundidas, tal como pueden encontrarse fácilmente en los textos más clásicos pero actualizados de psicometría. Así, podemos afirmar, en líneas generales, que **es posible distinguir cuatro grandes áreas en las que pueden categorizarse las distintas aristas e implicaciones del término *validez*: el área del contenido, él área empírica, el área teórica y el área formal.** De esta manera, haremos hincapié, con el fin de que nuestros alumnos comprendan más fácilmente este punto, en los *aspectos de contenido, en los aspectos empíricos o de aplicación, en los aspectos teóricos* y en los *aspectos formales de la validez* en las pruebas.

## 2.2 Aspectos de la validez vinculados con el contenido del test

Pasemos ahora al primer aspecto a ser considerado en los estudios de validez de una técnica: las evidencias de ***validez de contenido***. **Este tipo de evidencia de validez se refiere a la verificación de que la muestra de ítems incluida en el test cubra, efectivamente, todos los aspectos o dimensiones relevantes de la variable en estudio o a ser medida.** Para comprender este concepto, debemos pensar que *los reactivos o elementos que se han redactado o ideado para formar parte de un instrumento psicométrico son sólo una muestra de todos los ítems posibles, es decir, del universo de ítems destinados a medir esa variable en particular*. De todos los posibles, los finalmente incluidos (que no sólo han sido especialmente redactados sino que deben pasar por un estricto proceso de selección-eliminación; ver cap.5) deberán integrar una muestra representativa de todos los ítems posibles, guardando idénticos criterios de representatividad a los usados para resolver el muestreo de sujetos con los que se trabajará para estandarizar el test. Ello implica que no sólo deberán elegirse elementos claros y de buena calidad psicométrica (ver cap. 5) sino que los finalmente incluidos cumplan con el criterio fundamental de no haber descuidado ninguna de las dimensiones de la variable (García Cueto, 1993; Martínez Arias, 1999). Por ejemplo, si la teoría según la que se ha construido el instrumento postula que la variable Depresión se operacionaliza en tres dimensiones –por caso, ideación, componentes psicomotores y anhedonia– pero la primera dimensión posee, para este modelo, mayor importancia en la determinación de un cuadro depresivo, entonces los ítems a ser seleccionados e incorporados al test deberían aludir a los tres contenidos mencionados, habiendo una mayor cantidad de los vinculados a ideación, a fin de respetar el cuerpo teórico que basa la escala. Esta tarea debe ser desempeñada por los diseñadores originales del instrumento y posteriormente controlada en aquella instancia que se conoce como *juicio experto*: el juicio experto está destinado a trabajar sobre los aspectos de la validez que se relacionan con el contenido de los ítems –además de que también se la emplea, adicionalmente, como instancia de reformulación de consignas, giros gramaticales, jergas empleadas, estilos y formatos de respuesta y materiales, entre otras cuestiones a ajustarse antes de lanzar el instrumento al ámbito de aplicación–.

El juicio experto es un procedimiento mediante el que los autores del test convocan a un pequeño grupo de expertos en el tema que se quiere evaluar mediante esa nueva escala y, si es posible, siempre resulta preferible que tales expertos tengan, además, conocimientos de psicometría. Por ejemplo, si se quiere construir una herramienta destinada a medir la ansiedad según postula un modelo teórico dado, deberá convocarse a expertos en el tema –clínicos especializados en el tratamiento de trastornos de ansiedad, investigadores dedicados al estudio del concepto ansiedad y teóricos dedicados al análisis de la teoría mencionada– para que juzguen, en una primera aproximación, la adecuación de cada uno de los ítems que se han formulado para formar parte del instrumento, en relación a su coherencia o concordancia con alguna de las dimensiones propuestas por el autor en ese modelo teórico. Así, estos expertos, ayudados por una lista detallada de especificaciones que los autores le facilitarán, revisarán el contenido y la redacción de cada ítem y su adecuación con respecto a cada dimensión prevista por la teoría, efectuando críticas a algunos de ellos, mostrando su conformidad con respecto a otros, sugiriendo modificaciones en otros casos y alertando sobre aquellos aspectos o dimensiones de la variable en cuestión que estuvieren menos representados en el total de elementos o hubieren sido descuidados en aquéllas.

Cada experto trabajará en forma independiente y hará llegar su evaluación a los autores quienes, en un segundo momento, sopesarán el dictamen de cada uno de los expertos en base al conjunto de todos los dictámenes recogidos, efectuando sobre los ítems propuestos aquellas modificaciones que hayan sido sugeridas por la mayoría –o, al menos, por una parte importante del grupo experto–. Los criterios a partir de los cuales los autores modificarán, reemplazarán o eliminarán ítems deberán haber sido fijados de antemano y ser respetados a rajatabla, con el fin de evitar que la subjetividad de los creadores de la herramienta interfiera en demasía en este proceso.

Así, este procedimiento implica vigilar activamente que todas las dimensiones de la variable a ser medida hayan sido debidamente cubiertas, con el peso y la importancia que la teoría asigna a cada una de las mismas (Anastasi & Urbina, 1998); resáltese que en un modelo teórico, no necesariamente todas las dimensiones reciben la misma relevancia o importancia por parte del o los autores que formularon la teoría, por lo que en el test diseñado para medir esa variable, cada aspecto, entonces, deberá hallarse representado con una cantidad de ítems o puntuación diferencial de los mismos que resulte consistente con la importancia que se le dé en el modelo.

Frecuentemente se califica este procedimiento de juicio experto como un medio *subjetivo* de aportar evidencias de validez teórica, aunque debe destacarse que al incluirse varios jueces en él –frecuentemente en número impar, para "desempatar" en caso de desacuerdos– y al fijarse con anterioridad los márgenes según los que se aceptarán o se rechazarán las modificaciones (expresados mediante porcentajes o coeficientes específicos[1]), se asegura una notable disminución de la subjetividad individual de cada experto; sin embargo, volviendo a lo que antes afirmábamos, no puede nunca hablarse de subjetividad nula en la actividad científica.

Este procedimiento no es categorizado por la mayoría de los autores como un estudio de validez de constructo (que veremos en el apartado 2.1.3.) sino de contenido, aunque el segundo tipo de validez supone a la primera ya que el contenido aludido en los ítems dependerá del modelo teórico que se haya operacionalizado; de allí

1. El lector interesado en profundizar estos índices (coeficientes Aiken y otros) puede consultar Aiken (1999).



el parentesco cercano, casi indisoluble e inseparable en la comprensión de ambos tipos de estudios que aportan evidencias de validez.

## 2.3 Aspectos empíricos de la validez (aspectos de la validez vinculados al criterio)

Antes de desarrollar el concepto de validez empírica o de criterio, resulta conveniente aclarar que estas diferenciaciones en tipos no implican que deba determinarse una u otra clase de validez, sino que un investigador responsable debería trabajar sobre todos –o bien sobre la mayor cantidad posible– de aspectos de la validez, encarando el estudio de los mismos en forma empírica, tal como venimos detallando hasta aquí y como seguiremos desarrollando en las páginas siguientes. Será, así, el usuario idóneo y responsable quien exigirá que tales investigaciones se hallen concluidas y que sus resultados hayan aportado sólidas evidencias de validez de la mayor cantidad de tipos posibles, o de todos. Sin estos resultados –recordemos la definición general de validez– no nos encontraremos en condiciones de saber fehacientemente qué mide la técnica (constructo/variable y dimensiones) y cómo lo mide (según qué teoría e indicadores).

Pasemos ahora a definir la palabra *empírico/a*. El Diccionario de la Lengua Española (Real Academia Española, 2001) dice: ".... que se rige por la experiencia, perteneciente o relativo a la experiencia, fundado en ella" (p. 888). Esta definición implica dos cuestiones; la primera se relaciona con que este tipo de evidencias de validez, como las anteriores, deben ser establecidas empíricamente, es decir, en base a estudios de campo realizados con rigor metodológico y sobre personas concretas –aquí se aplica la parte de la definición antes leída: "que se rige por la experiencia, fundada en ella"–. En segundo lugar, el término *validez empírica* remite a un tipo de evidencia que se vincula con el uso práctico –como su mismo nombre lo indica, *empírico*– que puede darse al test, en base a los resultados que arroja, es decir, en base a qué mide y cómo mide en la práctica, en virtud de qué información nos aporta en concreto si lo administramos a un tipo de sujetos en particular. En última instancia, involucra una metodología de investigación que, necesariamente, trabajará utilizando lo que se conoce como *criterio externo*.

Es decir que un estudio orientado a aportar evidencias de validación empírica de los resultados arrojados por un instrumento deberá contar con un criterio externo, que es una medida de la misma variable que el instrumento intenta medir, pero obtenida en forma independiente de aquél. Será esa medición externa a la prueba, en este caso, la que nos permita verificar empíricamente si la escala nos brinda una información semejante a la obtenida mediante esa medición independiente, o criterio externo. Aquí se alude a la parte de la definición que reza: "perteneciente o relativa a la experiencia": se trata de un procedimiento netamente empírico, ya que apunta al uso práctico de la técnica, se realiza empíricamente (en este punto no se diferencia de las formas en que se trabaja para recoger evidencias de validez de constructo y de contenido) y no focaliza su atención en los aspectos teóricos (aunque tampoco los pierde de vista) sino en los empíricos. Decimos que no hace foco en la teoría, sin descuidarla, ya que el objetivo último de estos estudios es el de aportar resultados vinculados al uso empírico o práctico del test, pero por supuesto, el criterio externo que se emplee para llevar a cabo esta investigación deberá estar sustentado en el mismo marco teórico que el instrumento cuyos resultados se desean validar y, por ende, operacionalizado en forma similar.

## La validez concurrente

Reflexionemos ahora sobre la utilización práctica de una herramienta psicométrica en el ámbito de aplicación. ¿Qué busca en ella un evaluador que se desempeña en los ámbitos de evaluación forense, educativa, clínica, laboral o de programas? Sin dudas, recabar cierta información acerca de determinados atributos del sujeto en el menor tiempo posible, con el fin de responder determinadas preguntas relacionadas con un problema, interrogante o consulta específica (ver cap. 1). Pues entonces, para que estemos seguros de que un nuevo test, del que aún no conocemos su calidad, nos permite acceder a esa *cierta información* que necesitamos conocer, debemos poder corroborar que arroja *esa misma información o los mismos resultados que podríamos obtener por otros medios* (Anastasi & Urbina, 1998; Martínez Arias, 1995). Pongamos un ejemplo. Si un autor ha elaborado una prueba que mide el rendimiento en actividades intelectuales relacionadas con las tareas académicas clásicas que se plantean a nuestros estudiantes en el nivel medio de enseñanza, para verificar si estamos midiendo aquello que la escala pretende medir, además de asegurarnos de que los pasos relativos a las evidencias de validez teórica y de contenido hayan sido examinados, deberemos controlar también que se haya corroborado fehacientemente que puede medirse lo mismo –o casi lo mismo– que el test promete medir mediante un camino o criterio independiente al mismo test, externo a él. Es por ello que la validez empírica es conocida también como *validez de criterio* –ambas expresiones pueden emplearse como sinónimos–, ya que necesariamente implica el uso de un criterio externo que, como ya explicamos, se define como una medida independiente –diferente del instrumento– de la misma variable que quiere medirse. De esta manera, si el autor aporta evidencias de que puede medirse lo mismo –o casi– mediante ambas medidas –técnica y criterio externo– estará demostrando, esta vez desde una perspectiva empírica o pragmática, que su instrumento mide aquello que ha prometido en su denominación comercial o académica.

Veamos otro ejemplo: cuando un docente que trabaja en el nivel primario observa sistemáticamente a sus alumnos en la tarea escolar cotidiana y toma en cuenta cada una de las evaluaciones y trabajos encomendados, si está lo bastante atento, logra conocer a fondo la capacidad potencial de cada uno de sus estudiantes para lograr determinado rendimiento en un área dada. Esto le llevará un tiempo largo de observación y de sistematización de tanta información que se logrará sólo luego de mucho trabajo e interacción con cada educando. Ahora bien, si es posible diseñar una escala psicométrica que, apuntando a los contenidos escolares deseados, sea capaz de evaluar en un lapso breve y con poco esfuerzo la potencialidad de cada estudiante para acceder a determinado nivel en su rendimiento académico, entonces, hipotéticamente, este test abreviaría la tarea del docente ya que le brindaría una información comparable a la derivada de la observación en un lapso mucho menor y con menos esfuerzo. Sin embargo, para probar que esta nueva herramienta aporta los mismos resultados que la observación de cada alumno, será necesario efectuar una pequeña investigación que permita corroborar la equivalencia de tales resultados. ¿Cómo hacerlo?

Debemos aclarar que esta actividad será llevada a cabo por investigadores y no por el usuario o administrador quien, de todas maneras, deberá hallarse suficientemente capacitado para interpretar adecuadamente los resultados de las mencionadas evidencias de validez empírica, con el fin de decidir responsablemente el empleo de ese instrumento en tal o cual persona real y concreta.



Para ello, el investigador deberá administrar el test a una muestra de sujetos que reúnan características de sexo, edad, residencia y otras similares a las que se han especificado para las personas a las que está destinado el instrumento. Si se ha diseñado una escala para niños escolares de Buenos Aires, la muestra de sujetos sobre la que se trabajará estará compuesta por típicos niños escolares de Buenos Aires, evitando que dicha muestra se componga sólo de alumnos muy destacados o, por el contrario, de alumnos con dificultades. Se trabajará, así, con estudiantes de rendimiento medio o con todos los alumnos que se encuentren típicamente en la población general –de rendimiento alto, medio y bajo–. A la vez, a la misma muestra de sujetos a la que se ha aplicado la prueba se la hará objeto de una observación sumamente detallada y minuciosamente pautada –que funcionará como criterio externo–, que deberá ser determinada a priori con idénticos criterios que los utilizados para la construcción de la escala; ello significa que ambas vías de evaluación –técnica y criterio externo– apuntarán, al menos hipotéticamente, a medir la misma variable y según el mismo modelo teórico, ya que ambos han sido diseñados con tal propósito

Se empleará un coeficiente de correlación para valorar el grado en que ambos caminos de evaluación se hallan asociados, es decir, coinciden. Recordemos que el coeficiente de correlación es un índice que nos informa el grado de covariación o asociación entre dos variables, tratándose de una correlación directa (de signo positivo) cuando ambas variables aumentan o disminuyen juntas, y siendo la correlación inversa (de signo negativo) cuando al aumentar una de las variables, la otra disminuye o viceversa (Botella, León & San Martín, 1997). Advertimos ejemplos de correlación directa cuando nos referimos a la asociación entre la cantidad de trabajo que tenemos y la cantidad de tiempo que nos llevará terminarlo, cuando hablamos de la cantidad de harina que lleva una comida según la cantidad de personas que vayan a comerla, al nivel de rendimiento académico y el nivel de motivación para aprender, por ejemplo. A mayor cantidad de trabajo, más tiempo necesitaré para finalizar; a más personas invitadas a comer, más harina precisaré; a más motivación para aprender, mejor rendimiento escolar de los estudiantes. Otros ejemplos: en la infancia, la edad y la maduración conceptual; la motivación y el logro, la ansiedad y los bloqueos en situaciones ansiógenas.

Hablamos, en cambio, de correlaciones inversas o negativas cuando analizamos la relación entre la antigüedad del modelo de un auto y su precio –más antiguo el auto, menos se cotizará en el mercado–, la cantidad de trabajadores para realizar una tarea dada y el tiempo a emplear para finalizarla –a más trabajadores, menos tiempo de trabajo–, la importancia o gravedad de una depresión y la posibilidad de sentir interés o placer por ciertas actividades –a mayor depresión, menor interés–. Otros ejemplos: la cantidad de veces que se recurre a rituales compulsivos y el grado de salud mental de un sujeto, la gravedad de un trastorno psicótico y el nivel de adaptación a la vida laboral.

Independientemente del signo o sentido de la correlación, el grado de asociación entre las variables consideradas será mayor cuanto más se acerque su coeficiente a uno; será menor cuanto más cerca de cero se ubique. Así, una correlación perfecta de +1 nos indicará una covariación directa entre las variables (ambas aumentan o disminuyen juntas) y perfecta (las variables están asociadas completamente, en un 100%); (Botella, León & San Martín, 1997). Un ejemplo de correlación perfecta puede estar dado por la cantidad de tazas que necesito para convidar con café a mis invitados si todos ellos toman esta infusión. Necesitaré una taza por invitado. Al aumentar la cantidad de invitados, aumentará la cantidad de tazas requeridas y ambas magnitudes se

incrementarán al mismo ritmo: por cada invitado extra que se añada, se agregará una taza más; la correlación, así, será perfecta, es decir, igual a 1.

Una correlación será nula (igual a cero) cuando la asociación entre las variables sea inexistente. Por ejemplo, la cantidad de dinero que alguien tiene en el banco y la cantidad de lunares que advierte tener en su cuerpo.

Tenemos correlaciones no perfectas cuando encontramos asociación entre dos variables - cualquiera sea su signo - distintas de cero y distintas de uno (es decir, no perfectas ni nulas). Allí hallaremos coeficientes de correlación de, por ejemplo, 0,15 ó 0,90. Una correlación de +0,15 (ó de -0,15) implica un grado bajo de asociación entre las variables –más cercano a cero– mientras que +0,90 (ó -0,90) muestra un valor cercano a 1 y, por ende, elevado; este último indica un alto grado de covariación entre las variables consideradas. De ninguna forma una correlación –aún una muy elevada– puede interpretarse como una variable causando o determinando a la otra, sino que simplemente se muestra la asociación o relación entre ellas, siendo las hipótesis causales terreno de inferencias teóricas que exceden al coeficiente de correlación.

En el caso de un estudio de validez concurrente, si este coeficiente es positivo y elevado, estará indicando que ambas mediciones –criterio y escala– realizadas a una única muestra de sujetos arrojan resultados muy similares, por lo que será prácticamente lo mismo medir la variable deseada mediante la prueba o mediante el criterio –observación de tareas escolares–. En el uso práctico en el ámbito de aplicación que se dará con posterioridad al proceso de validación y como resultado de él, si se dispone de tiempo podrá utilizarse la segunda vía, en tanto que si se desea conocer estos resultados en un lapso muy breve y con menos esfuerzo, se administrará el test. Así, aquel estudio ha arrojado **evidencias de validez empírica o de criterio, de tipo *concurrente*.** Esta expresión significa que técnica psicométrica y criterio *concurren* juntos, en un mismo sentido, arrojando idénticos resultados –o muy similares–; los textos clásicos de psicometría dirán que la información obtenida mediante el instrumento es reemplazable o intercambiable por la información brindada por el criterio, que el instrumento se propone como sustituto de otro tipo de información (que, en este caso, será el criterio); (Anastasi & Urbina, 1998; Casullo, Figueroa & Aszkenazi, 1991). Sencillamente expresado, es lo mismo –o casi lo mismo– administrar la prueba u observar a los alumnos en su tarea diaria durante cierto intervalo, ya que ambos procedimientos nos permitirán acceder a información similar. Por supuesto que prueba y criterio deberán estar operacionalizados según la misma base teórica, tal como decíamos antes. Supongamos que leemos en el manual de una herramienta psicométrica que:

> "…..se administró una escala para evaluar impulsividad a una muestra de 250 sujetos de población general, residentes en tal ciudad, cuyas edades variaban de 18 a 35 años; a su vez, se efectuaron, con la misma muestra, en los cinco días subsiguientes, entrevistas individuales a cargo de dos psicólogos clínicos que debían arribar a un diagnóstico positivo o negativo de presencia y nivel de impulsividad a partir de las mismas –utilizadas como criterio externo–. En ellas se trabajó con una grilla de puntuaciones que permitió obtener una puntuación global para cada sujeto, derivada de las puntuaciones parciales relacionadas con los signos y síntomas posibles previstos para diversos comportamientos impulsivos. Esta puntuación global se correlacionó con la derivada del mencionado test de impulsividad, obteniéndose un coeficiente de correlación de 0,97; $p<0,01$, aportándose, así, contundentes evidencias de validez concurrente entre el test y el criterio externo".



El párrafo anterior, que podríamos hallar en un hipotético manual de un hipotético test, implica que en sujetos de población general, residentes en tal ciudad, con edades entre los 18 y los 35 años, el instrumento mencionado permite arribar a un diagnóstico dado de impulsividad equivalente al de una entrevista individual con un muy elevado grado de certeza.

El ejemplo anterior nos indica que el test y la entrevista pueden ser utilizados en forma indistinta ya que ambos arrojan información muy similar, pero a la vez nos muestra que el instrumento evalúa, efectivamente, impulsividad –o comportamientos impulsivos, según el autor de la prueba lo hubiere delimitado–, de manera coherente a como ha sido planteado en la entrevista –criterio externo–, ya que el coeficiente de correlación es cercano a uno (0,97; la asociación entre ambos resultados es casi total) y significativo ($p<0,01$); expresado de manera muy simplificada y sin mayores pretensiones estadísticas, este p valor inferior a 0,01 quiere decir que existe al menos un 99% de probabilidades de que ese coeficiente de correlación obtenido se deba a la influencia de las variables que se han correlacionado (puntuaciones de los sujetos en el test y puntuaciones de los mismos sujetos en las entrevistas) y sólo existe un 1% o menos de probabilidad de que esta correlación se deba al azar. Por otro lado, cuando más cercano a 1 sea el coeficiente de correlación obtenido, mayor será la concurrencia o coincidencia entre las puntuaciones arrojadas por el test y el criterio externo; de allí surgen las denominaciones de validez empírica concurrente (concurrencia empírica entre test y criterio externo) y de validez de criterio (en base al ya mencionado uso de un criterio externo en la investigación diseñada para verificarla). Ello significará que ambas vías para medir la misma variable aportan información semejante o muy semejante, según sea el valor del índice de correlación. Y de esta forma, se aportan evidencias acerca de la validez concurrente de los resultados obtenidos mediante un test dado, de manera que éste se propone como sustituto de la información aportada por el criterio (pues ambos miden casi lo mismo), verificándose la posibilidad de utilizar en la práctica el instrumento en una población homogénea (con las mismas características) con respecto a la muestra sobre la que se efectuó este estudio. De allí su nombre de validez empírica.

Ejemplos de criterios externos a ser utilizados en el ámbito de la psicología podrían ser:

- Calificaciones académicas para aportar evidencias de validez de los resultados de una prueba de rendimiento en alguna asignatura determinada en el nivel universitario de enseñanza.
- Una subescala de un test de personalidad que mida depresión para brindar evidencias de validez de los resultados de una técnica que mida ese mismo constructo en pacientes adultos.
- El rendimiento real observado en tareas visomotrices de coordinación ojo-mano para recabar evidencias de validez de los resultados aportados por un test de maduración visomotriz en escolares.

Así, como puede verse en los ejemplos anteriores y como hemos afirmado varios párrafos atrás, un criterio externo es una medida de la variable, establecida en forma independiente al instrumento (externa a él), cuyas evidencias de validez, a su vez deberán estar previamente establecidas, además de ser coherentes con la base teórica que sustenta la prueba puesto que, de otra forma, no serían comparables. Es

por ello que muchos manuales tradicionales de psicometría introducen como nomenclatura técnica de validez empírica la designación $r_{xy}$, puesto que $r$ simboliza al coeficiente de correlación $r$ de Pearson —, $x$ representa las puntuaciones obtenidas en el test por la muestra de sujetos, en tanto que $y$ representa la puntuación obtenida por esos mismos sujetos en el criterio externo. Debe señalarse que aunque existen otros coeficientes de correlación, el más usual en el tipo de estudios que aquí se comenta es el $r$ de Pearson.

Para diferenciar claramente los conceptos de validez y confiabilidad (ver Cap. 4), podemos adelantar que este último concepto se representa mediante la nomenclatura $r_{xx}$, donde $r$ simboliza también al r de Pearson, $x$ a las puntuaciones obtenidas en la prueba por la muestra de sujetos sobre la que se trabajó, y la segunda $x$ representará a las puntuaciones obtenidas por la misma muestra de sujetos en una segunda aplicación del instrumento, o a las derivadas de una forma paralela del test, administrada también a los mismos sujetos. Sin ahondar en detalles que se desarrollarán en el Cap. 4, sólo diremos, a partir de las nomenclaturas antes analizadas, que *mientras la validez empírica ($r_{xy}$) se dirige a estudiar la relación entre las puntuaciones arrojadas por el instrumento y las obtenidas mediante el criterio externo*, la *confiabilidad ($r_{xx}$) analiza las puntuaciones al interior de la prueba o las puntuaciones de ésta comparadas con las de una forma paralela de la misma*, que es prácticamente lo mismo que referirse al test original –expresado en forma más que simplificada–. La validez empírica siempre trabaja en forma externa a la técnica (criterio externo), en tanto que la confiabilidad lo hace en forma interna (examinando la consistencia entre las puntuaciones obtenidas mediante la escala *al interior del conjunto de los ítems que la componen, o comparando contra una forma paralela* del test, o bien comparando las puntuaciones obtenidas en dos administraciones sucesivas de la técnica, que es casi lo mismo que decir que se compara a la prueba consigo misma). Ampliaremos estas nociones en el Cap. 4

Ahora bien, cuando aludimos al uso de un criterio externo en los estudios de validez concurrente, inmediatamente surge una pregunta práctica: si test y criterio brindan información comparable, ¿para qué usar el primero, si con el segundo ya basta? Antes que nada, tengamos presente que este interrogante se deriva del uso concreto de la técnica con individuos reales y concretos, en el ámbito de aplicación y no ya en el de investigación. La formulación o elección de un criterio externo se relaciona con necesidades inherentes al diseño de investigación que se plantea en un estudio que pretende aportar evidencias de validez empírica de los resultados a los que el instrumento conduce, pero ello no es obstáculo para que en el ámbito de aplicación el usuario elija libremente aquella vía de evaluación que prefiera, en base a su modalidad de trabajo, a los tiempos y condiciones disponibles y al esfuerzo y cantidad de información resultantes que está en condiciones de invertir y de procesar. Si comparamos una observación de rendimiento con un test de rendimiento, está claro que la primera permitirá más información sobre menos casos, en tanto que el segundo brindará información menos profunda pero sobre más casos y en menos tiempo. Es aquí donde el usuario que se desempeña en el ámbito de aplicación deberá tomar decisiones, y si dispone de dos o más vías alternativas, tanto más rico y amplio será su margen de opciones. A la vez, esa disponibilidad le permitirá escoger la medición más adecuada a las características de los sujetos con los que trabajará, así como al motivo de la evaluación o a la inquietud a la que se desea responder.



## La validez predictiva

Siguiendo con los tipos de evidencias de validez que aún nos resta revisar, introduzcámonos ahora en otra variante de la validez empírica o de criterio, la ***validez predictiva***. Comparte con la validez concurrente su pertenencia a la categoría de validez empírica o de criterio; ambas son los dos tipos principales previstos en ella. Posee, además, una lógica similar a la de la validez concurrente trabajando, asimismo, con un criterio externo y un coeficiente de correlación. La diferencia es que la validez concurrente se establece en y para el momento presente –aquí y ahora– con el fin de asegurar que la escala mide aquella variable que prometía medir: prueba y criterio externo deberían medir la misma variable en el aquí y ahora para que se concluya que se han aportado evidencias de validez concurrente. La validez predictiva, en cambio, trabaja a futuro y con un criterio externo a predecirse, diferente de la variable medida en el aquí y ahora por el test, también empleando el coeficiente de correlación como medida de las evidencias de validez aportadas (Anastasi & Urbina, 1998). Aclaremos un poco más este punto.

Un estudio destinado a aportar evidencias de validez predictiva se basa en la idea de intentar verificar que el instrumento –administrado en el presente– resulte un buen predictor de otra variable –relacionada teóricamente con la que se ha medido, pero distinta– cuyo comportamiento futuro interesa estimar (Casullo, Figueroa & Aszkenazi, 1991).

Por ejemplo, se planifica usar una prueba de capacidad de aprendizaje de contenidos verbales que se administra a los ingresantes a primer año de la escuela media, con el objeto de predecir el rendimiento futuro de estos alumnos en materias en las que predominen los contenidos verbales, a lo largo del ciclo básico de la escuela secundaria (1° a 3er. curso). Para ello se vuelve indispensable lograr demostrar que el test de capacidad de aprendizaje de contenidos verbales administrado hoy resulta un buen predictor del desempeño futuro del alumno en las materias más impregnadas de contenidos verbales. ¿Cómo hacerlo?

Se administrará la prueba mencionada a una muestra representativa de ingresantes a primer año del ciclo medio a principios del año escolar. Los protocolos producidos por ellos se guardarán durante los tres ciclos lectivos que el estudiante atravesará luego de la administración del test – 1°, 2° y 3°– y cuando estos alumnos concluyan el tercer año de su escuela media, se obtendrá el promedio de sus calificaciones en aquellas asignaturas en las que predominen aquellos contenidos verbales en cada año cursado, que eran objeto de interés en este caso. Estas calificaciones obtenidas harán el papel de criterio externo que, en este ejemplo es aquello que se quiere predecir –el rendimiento en materias verbales– y se las correlacionará con las puntuaciones obtenidas por la muestra de alumnos en el test de capacidad de aprendizaje verbal administrado tres años antes. Si las correlaciones son altas (lo más cercanas a 1 posible) y positivas, entonces podrá concluirse que el mencionado instrumento, administrado a principio del primer año del ciclo medio en este tipo de alumnos, resulta un buen predictor del rendimiento en asignaturas verbales durante el ciclo básico de ese nivel de enseñanza. Expresado en forma muy simplificada, para que tal correlación sea elevada y positiva, los resultados obtenidos en la técnica deberían coincidir en mucho con los del criterio externo (rendimiento medido según calificaciones), en la mayoría de los casos analizados. Así, mayoritariamente, quienes hayan tenido un buen resultado en el test de capacidad de aprendizaje verbal registrarán también un buen

rendimiento académico en esas materias; quienes hubiesen mostrado un desempeño medio en la prueba, deberían exhibir, en líneas generales, un rendimiento coherente en las asignaturas verbales y aquellos con bajos puntajes en el instrumento psicométrico deberían haber obtenido, en su mayoría, bajas calificaciones en el área verbal de su boletín escolar a futuro.

En el ejemplo anterior puede inferirse que se trata del siguiente esquema: se administra el test para medir una variable *x* en el presente, con el objeto de verificar si sus puntuaciones son capaces de predecir a futuro el criterio (la variable *y*, diferente de la variable *x* medida por el test, pero relacionada con ésta en forma teórica y empírica). Dado que usa también un criterio externo en su procedimiento, comparte con la validez concurrente la pertenencia a la categoría validez de criterio; la diferencia es que en la predictiva el criterio se predice y es ***otra*** variable, diferente de la medida mediante el test, que se mide a futuro[2]. En la concurrente, en cambio, el criterio se mide en el mismo momento –o casi– en que se administra la prueba, y es ***la misma*** variable evaluada por el instrumento, pero obtenida por medio de una medición independiente a la prueba. Y, lógicamente, el rótulo de predictiva obedece a que se intenta predecir el criterio a futuro, en tanto que en la concurrente se intenta que concurra o coincida con los resultados de la escala en el mismo corte temporal. Sin embargo, debe notarse que, tal como aseveran Anastasi y Urbina (1998) la diferencia lógica entre validación concurrente y predictiva no se basa en el tiempo sino en los objetivos de la evaluación: mientras que la primera es la elegida cuando la prueba va a emplearse para efectuar diagnósticos del estado actual, la segunda lo será cuando el instrumento busque predecir resultados futuros; es decir que el uso de la escala en el ámbito de aplicación o de investigación será el criterio decisorio que permitirá dirimir qué tipo de estudio de validación se empleará.

La validez predictiva se clasifica dentro de la validez empírica, junto con la concurrente, porque ambas hacen al uso práctico o empírico de la técnica, a su utilidad. Un usuario que lee en el manual de tal técnica que se ha realizado un estudio que aporta evidencias de validez predictiva como el que se comenta párrafos atrás, puede asumir que, empleado con sujetos semejantes a los que participaron del estudio y bajo idénticas condiciones, el test será una herramienta útil para la predicción de la variable –criterio en el ámbito de aplicación y como tal, podrá hacer uso de ella en su quehacer cotidiano–. Es decir que, una vez que en el ámbito de investigación se haya arribado a determinados resultados positivos sobre la capacidad de un instrumento para predecir ciertos comportamientos, en el ámbito de aplicación se tomarán en cuenta esos estudios para estimar –con cierto margen de certeza– que tal individuo en particular que ha sido evaluado mediante esa escala, *probablemente* presentará también ese comportamiento, sin necesidad de tener que replicar ese estudio.

Un concepto a ser tenido en cuenta en el tema de las evidencias de validez predictiva es de *validez incremental*. Es un hecho bastante frecuente que en diversos ámbitos de aplicación los evaluadores tengan interés en establecer la probabilidad de predecir algún criterio (o comportamiento o rendimiento dado) a partir de una multiplicidad de tests y no de uno solo o, dicho de otro modo, a partir de varios predictores y no de un único predictor. Cada test empleado como predictor debería, por supuesto, contar con evidencias sobre su validez predictiva en forma independiente, vinculada con el criterio a predecir. El concepto de validez incremental, entonces, implica conocer el grado en que cada predictor –cada test usado para predecir el criterio– explica o predice algo

2. Aunque algunos autores cuestionan que necesariamente deba hablarse de predicción "a futuro", pero profundizar en este aspecto iría mucho más allá de las pretensiones de este libro.



de la medida del criterio que no estaba predicho por los otros tests o predictores. Por ejemplo, el rendimiento académico podría predecirse a partir de resultados obtenidos en varios tests: uno de razonamiento verbal, uno de razonamiento abstracto, otro de estrategias de aprendizaje, otro de motivación y otro de ansiedad ante los exámenes. Si sólo intentamos predecir el rendimiento académico mediante el desempeño en el test de razonamiento verbal, el criterio podrá ser estimado con un margen de probabilidad dado. Si se agrega un segundo predictor (razonamiento abstracto), se *incrementaría* el grado en que se predice el criterio rendimiento académico; sucesivamente, si se añade, en cada paso, un nuevo predictor, se incrementa la capacidad de ese grupo de predictores (tests) para estimar el criterio (rendimiento académico); en ese postulado básico se asienta, entonces, el concepto de validez incremental.

### La validez retrospectiva

A pesar de que no es un aspecto muy comúnmente tomado en cuenta, resulta de interés referirse a la *validez retrospectiva*, considerada por García Cueto (1993) como un aspecto de la validez empírica o de criterio, junto con la concurrente y la predictiva. Se vincula con la correlación verificada entre los resultados de un test administrado en un momento determinado y un criterio externo medido con antelación a la aplicación del instrumento psicométrico, aún años antes. Según este autor, este tipo de evidencias de validez adquieren importancia en ciertos ámbitos específicos, tales como la prevención psicopatológica en la salud pública o en la clínica individual, por ejemplo.

### Otros estudios posibles

Otra manera posible para examinar la validez de los resultados obtenidos mediante un test es efectuar un estudio por grupos contrastados. En estos casos, el objetivo consiste en demostrar que las puntuaciones arrojadas adquieren valores predecibles en función de la pertenencia de los individuos a un grupo dado (Cohen & Swerdlik, 2000). Cuando un criterio determinado ha sido establecido de antemano –por ejemplo, ya se cuenta con un grupo de pacientes que han sido diagnosticados como presentando ideación suicida– pueden validarse los resultados aportados por una escala que evalúe esa variable, de manera que el grupo de pacientes con tal diagnóstico deberían puntuar significativamente más alto que otro grupo de no-pacientes, utilizados en esta hipotética investigación como grupo de comparación. Si la escala está midiendo correctamente el constructo que se ha propuesto, debería ser capaz de discriminar fácilmente quiénes son aquellas personas que presentan esta clase de pensamientos y quiénes no; de esta forma, puede advertirse cómo este tipo de diseños de investigación contribuye a aportar evidencias acerca de la validez de constructo de una prueba, aunque también, indirectamente, brinda evidencias de validez empírica.

Análogamente, podrían también diseñarse estudios con grupos contrastados destinados a determinar evidencias de validez predictiva; por caso, sería factible, para validar la capacidad predictora de los resultados arrojados en el presente, dado un test que evalúe capacidad potencial para desempeñarse en tareas administrativas en empresas de envergadura, administrado al momento de la recepción del currículum vitae de los postulantes. Una vez asignados todos los candidatos en diversos puestos de trabajo de naturaleza y dificultad similar, a los seis meses del ingreso –o en algún

plazo determinado–, se procederá a evaluar su desempeño real como empleado administrativo en tales empresas. En base a estas evaluaciones (criterio a predecirse), se separará a los empleados en dos grupos: empleados eficientes y empleados que deben mejorar. Una vez segmentada la muestra de sujetos en estos dos grupos, se traerán a colación los protocolos del test de potencial de desempeño, administrado seis meses antes. Si las puntuaciones obtenidas en ese instrumento guardan diferencias estadísticamente significativas, de manera que los que resultaron empleados eficientes en la realidad lograron desempeños significativamente mejores en el test y los que mostraron trabajar en forma menos adecuada habían puntuado en la franja de desempeños sensiblemente menores en la prueba administrada un semestre antes, ello significa que el instrumento resulta un buen predictor del comportamiento o criterio a predecirse. Este estudio, desarrollado tal como se ha detallado hasta aquí, servirá de base para hipotetizar que la prueba, administrada a candidatos que reúnan características homogéneas a los incluidos en la muestra analizada, será un buen predictor del desempeño real ulterior, de manera que, en el futuro, y ya trasladándonos al campo de aplicación, solamente administrando la técnica se podrá estimar –con cierto margen de certeza– la capacidad de un candidato para desempeñarse en el puesto que solicita, sin necesidad de aguardar a evaluar su rendimiento posterior en el rol. Estas conclusiones servirán como criterio de decisión a los evaluadores que se desempeñen en el ámbito de la evaluación laboral, a los fines de emplear esta herramienta que ha logrado acumular evidencias sobre la capacidad predictiva de sus resultados. Por ende, bajo determinadas condiciones y con cierto margen de certeza calculable, se puede suponer que los resultados obtenidos en esta técnica por un candidato al puesto, permitirán predecir su rendimiento real ulterior en esa posición laboral. Debe aclararse que, por supuesto, la investigación que da soporte a tales conclusiones, exige que *todos* los postulantes sean admitidos (evento, de por sí muy complicado y bastante poco económico para una organización) y evaluados seis meses más tarde, pero tal operación no debería repetirse al momento de emplear la prueba en el campo de aplicación laboral, sino que se confiará en las conclusiones informadas, siempre y cuando se cumplan determinadas condiciones de rigurosidad científica y de semejanza entre los miembros de la muestra empleada y los candidatos.

## 2.4 Aspectos de la validez vinculados con el modelo teórico que sustenta la prueba

Comenzando por los aspectos teóricos, debe recordarse, tal como afirmábamos en el Cap. 1, que toda escala psicométrica no es, desde el punto de vista metodológico, ni más ni menos que la operacionalización de un constructo, en la manera como está previsto en un modelo teórico dado. Por lo tanto, diremos que **tales aspectos teóricos de la validez se circunscriben a que los autores del test u otros investigadores sean capaces de aportar evidencias de que tal operacionalización ha sido efectuada en forma coherente con ese modelo teórico y cubriendo todos los aspectos o dimensiones incluidos en él.** Desmenucemos mejor esta aseveración.

Por un lado, los investigadores especializados en psicometría que se han ocupado de diseñar o de estudiar un test construido por otro investigador, deberán aportar evidencias verificables de que ese instrumento mide efectivamente el constructo o variable que dice medir, es decir, que tal constructo ha sido adecuadamente operacionalizado en indicadores (ítems del test) capaces de aportar mediciones adecuadas de los



distintos aspectos o dimensiones de tal variable. Como afirmábamos en el primer apartado de este capítulo, no debemos tomar como palabra sagrada e incuestionable el nombre de una escala, la cual generalmente indica qué variable (supuestamente) es medida por medio de tal herramienta: en cambio, debemos ahondar más en la cuestión, leyendo concienzudamente el manual de la técnica, y analizando con minuciosidad el capítulo referente a las evidencias de validez recogidas por distintos autores en diversas muestras de sujetos. Para muchos autores, **el tipo principal de evidencia de validez, la teórica –también llamada estructural o de constructo–, se dedicará, precisamente, a responder a la pregunta de si esta técnica mide efectivamente aquello que dice medir, según tal o cual modelo teórico y, por lo tanto, si la misma es una adecuada operacionalización de un constructo teórico dado, derivado de ese modelo** (Casullo, Figueroa & Aszkenazi, 1991) (ver Cap. 1).

Ahora bien, ¿cómo recoger este tipo de evidencias de validez? ¿Cómo establecer en forma no subjetiva –o, lo menos subjetiva posible– si este test es una adecuada operacionalización del constructo o no?

Ante todo, dos aclaraciones. La primera se dirige a establecer que tales estudios de validez no serán desarrollados por los usuarios que emplean las técnicas en el ámbito de aplicación –clínico, laboral, educativo, forense y de evaluación de programas (ver cap. 1)–, sino que serán los propios autores del test, u otros investigadores interesados en él como objeto de estudio (ver Cap. 1)- quienes llevarán a cabo tales investigaciones, pero será competencia y obligación de todo buen usuario o aplicador de las técnicas leer, analizar y juzgar la pertinencia y corrección metodológica de las mismas. Ello significa que será el usuario quien, a la luz de la lectura de esos resultados, decidirá emplear o no esa prueba para tales o cuales fines, ya que esas evidencias de validez son uno de los *certificados* que permiten juzgar la calidad del instrumento. Y, como todos sabemos a partir de nuestra experiencia de vida como consumidores, existen productos de buena, regular y mala calidad; y una prueba no es ni más ni menos que un producto tecnológico.

La segunda aclaración se vincula con destacar que cuando nos preguntamos cómo efectuar una valoración objetiva de si el test es una adecuada operacionalización de un constructo, nos estamos refiriendo a trabajar con la menor subjetividad posible, ya que jamás es factible despojarse de subjetividades, aún en ciencia. Cuando se dice que el conocimiento científico es objetivo, se alude a que es metódico, verificable, comunicable, replicable y *lo menos sujetivo posible*. Toda actividad y todo conocimiento están teñidos, en mayor o menor medida, de algún grado de subjetividad. En ciencia intentamos reducirla a su mínima expresión, pero jamás se iguala a un valor nulo: siempre estará presente. En tanto seamos conscientes de ello, menos serán nuestros errores en ese terreno. Tomando en cuenta estas salvedades, **la validación de constructo se define como un proceso continuo –durable en el tiempo, en tanto requiere de investigaciones desarrolladas y renovadas en forma permanente– por medio del que se realizan múltiples investigaciones con el fin de poner a prueba diferentes hipótesis sobre la estructura interna del constructo, así como de sus relaciones con otras variables o constructos** (Martínez Arias, 1995). **También se la puede definir como el grado en que un test mide un constructo, en tanto es una buena operacionalización del mismo** (Casullo, Figueroa & Aszkenazi, 1991). Es importante recordar que la validación de constructo implica la acumulación gradual de diversas fuentes de información (Anastasi & Urbina, 1998), por lo que no es posible considerarla como un proceso terminado, sino que necesariamente requerirá de constante actualización e investigación.

Procedimientos más frecuentes para aportar evidencias de validez de constructo

Volvamos ahora a cómo recoger –y por ende, valorar– evidencias de validez teórica de los resultados de un test. ¿Cómo verificar que el mismo es una operacionalización metodológicamente bien realizada, y coherente con el modelo teórico? Existen, fundamentalmente, varias formas de hacerlo: estudios evolutivos y clínicos, el análisis factorial, la evidencia convergente y discriminante, la investigación meta-analítica, evidencias de cambio pretest-postest, las matrices multi-método/multi-rasgo, entre otras posibles (Cohen & Swerdlik, 2001; García Cueto, 1993; Martínez Arias, 1995). Examinemos cada uno de estos procedimientos.

a) *Estudios evolutivos*

Tomemos, por ejemplo, los *estudios evolutivos*. Si la una teoría y/o resultados empíricos de distintas investigaciones postulan que, por ejemplo, a medida que un niño crece y madura, se acrecienta su capacidad para efectuar algún tipo de actividad dada, un test que intente medir esa capacidad, debería corroborar esto mediante sus resultados. Por ejemplo, si se ha verificado en diversos estudios que los niños aumentan, junto con la edad cronológica, su habilidad de coordinación visomotriz en tareas de dibujo y reproducción de formas gráficas (por ejemplo, figuras geométricas o, incluso, letras), un test que hubiera sido construido para evaluar esta habilidad visomotriz debería, también, demostrar mediante sus resultados que, a medida que la edad aumenta, se incrementa tal habilidad en los niños. Este tipo de investigaciones evolutivas es uno de los métodos más sencillos para aportar evidencias sobre la validez teórica de un test.

b) *Estudios clínicos*

Con una lógica similar de razonamiento se procede cuando se usan *estudios clínicos* si el constructo a ser evaluado implica algún tipo de patología. Por ejemplo, los resultados aportados por el instrumento en pacientes psicóticos deberían ser significativamente distintos de aquellos obtenidos por sujetos no psicóticos (véase estudios por grupos contrastados, *en el apartado 2.3*). Seguramente, a estas alturas, el lector atento estará preguntándose por las diferencias entre el procedimiento que aquí se describe y el esbozado en el apartado indicado (*2.3*), que se relaciona con las evidencias de validez empírica. Si ello está sucediendo, significa que esta confusión está llevándolo a interrogantes cada vez más sutiles, ya que los diferentes aspectos de la validez se relacionan y, en muchos casos –como, por ejemplo, en éste–, un mismo tipo de estudio podría brindar evidencias hacia ambas aristas del tema de la validación de los resultados aportados por el instrumento. Todo uso empírico (validez empírica) se vincula con el modelo teórico (validez de constructo) y, a la inversa, todo resultado empírico que se obtenga impacta directamente en los cuestionamientos, refutaciones o confirmaciones que puedan hacerse sobre un modelo teórico.

c) *Análisis Factorial*

Pero pasemos, ahora, al *análisis factorial o estudios factoriales*, que son el tipo de investigación más comúnmente empleados en nuestros días para aportar evidencias sobre la validez de constructo de los resultados brindados por un instrumento (Cohen & Swerdlik, 2001; García Cueto, 1993; Martínez Arias, 1995). El análisis factorial es un procedimiento de análisis multivariante de los datos que permite –digámoslo de manera casi coloquial– analizar la variable, tal como ha sido medida por medio de un test en una muestra de sujetos dada, determinando qué dimensiones podrían aislarse en la misma. El *análisis factorial es definido como un método de reducción de datos* (Hair, Anderson, Tatham & Black, 1999), por el que es posible –como ese rótulo lo indica– disminuir la cantidad de datos a ser analizados o tenidos en cuenta. Por ejemplo, si hemos trabajado con una muestra de 500 sujetos, a los que hemos administrado una escala de 100 ítems, tenemos un total de 50000 respuestas a ser analizadas (100 ítems x 500 sujetos= 50000 respuestas o datos). Como es imposible para un ser humano analizar conjuntamente semejante conglomerado de información, se recurre a un método que permite reducir esos 50000 datos a una cantidad significativamente menor: unos pocos *factores* –de allí el nombre de análisis *factorial*– o *variables latentes*, que podríamos tomar como lo que ya conocemos con el marbete de *dimensiones de la variable*.



El análisis factorial se maneja por medio del cálculo de múltiples coeficientes de correlación entre las respuestas de todos los sujetos incluidos en la muestra a cada uno de los ítems del test. Así, se calcula la correlación de las respuestas dadas por el grupo al ítem 1 con cada total de respuestas dadas a cada ítem de la escala. Esto dará lugar a lo que conocemos como una matriz de correlaciones, en la que podremos examinar la correlación existente entre las respuestas dadas a cada par de ítems que podamos aislar en el test (se correlacionan las respuestas de cada ítem con las de cada uno del resto de los reactivos; todos los elementos con todos los ítems).

Este cálculo de múltiples coeficientes de correlación entre todas las combinaciones posibles de respuestas emitidas por los sujetos incluidos en la muestra implica, en última instancia, intentar conocer qué grado de asociación, relación o covariación tiene el total de las respuestas de todos los sujetos en todos los ítems de la escala; es decir, cómo se asocian, los contenidos a los que ellos aluden. Por supuesto que la correlación calculada no nos informará sobre los contenidos presentes en las formulaciones de los reactivos, pero sí nos hablará de su asociación y, de alguna manera, de su *semejanza*. Por ejemplo, un reactivo que interroga sobre la asiduidad con que una persona verifica si cerró la puerta de su casa o la llave del gas parece tener un contenido bien diferente de otro que pregunta sobre pensamientos recurrentes y no controlados por la persona. Estas dos respuestas, distintas en apariencia, pueden agruparsé bajo un coeficiente de correlación alto, ya que ambas forman parte del conglomerado sintomatológico de un trastorno obsesivo-compulsivo y, por lo tanto, es muy factible que una gran parte de los individuos que respondieron afirmativamente al primero hayan respondido también afirmativamente al segundo. Así, el análisis factorial detecta, según el grado de asociación entre las respuestas a los ítems, si dos reactivos distintos guardan entre sí algún grado de asociación (alto, medio, bajo y positivo o negativo). De esta forma, aquellos elementos que hayan registrado entre sí elevados grados de asociación se agruparán bajo lo que llamamos un *factor, dimensión o variable latente*. Cada uno de estos factores será un grupo de ítems que aludan a contenidos relacionados (no necesariamente parecidos o iguales, sino emparentados de alguna manera específica), cuya vinculación o asociación se ha verificado mediante la agrupación de aquellos reactivos que guarden considerable asociación entre sí (correlaciones elevadas). Luego será tarea de los investigadores determinar cómo podrá etiquetarse o nombrarse a cada factor, mediante el minucioso análisis de los contenidos aludidos en cada uno de los ítems que se han agrupado en ellos (Hair, Anderson, Tatham & Black, 1999; Martínez Arias, 1999). Por ejemplo, un investigador nombrará a un factor como *Depresión* cuando se incluyan en él reactivos asociados (con altos coeficientes de correlación entre sí, y por lo tanto que se han agrupado en un mismo factor) que se refieran a signos y síntomas presentes en cuadros depresivos.

**Si la cantidad de factores aislados o identificados y sus contenidos coinciden con la cantidad e identificación de las dimensiones previstas en la teoría, entonces, podrá decirse que el test es una adecuada operacionalización de tal marco teórico, o bien que mide tal variable según tal teoría x.** Es de esta forma como se efectúa uno de los procedimientos posibles para "vigilar" la validez teórica de los resultados aportados por un instrumento. Si el análisis factorial aísla un número distinto de dimensiones que las previstas en el modelo, o el mismo número pero con distintos contenidos, entonces deberá revisarse el instrumento a fin de detectar errores técnicos, teóricos y/o metodológicos. Si todo ello se descarta, podrá pensarse que es la teoría la que debe revisarse, al menos en la población estudiada en este caso puntual y, por supuesto, deberán replicarse los estudios y diseñarse otros nuevos para recoger más evidencia relacionada (Cohen & Swerdlik, 2001).

Conviene agregar que la metodología del análisis factorial es un procedimiento controvertido, ya que existen, potencialmente, infinitas soluciones factoriales capaces de explicar un conjunto de datos determinado; la elección de una u otra estará sujeta a una serie de criterios técnicos pero dependerá, en última instancia, de la metodología empleada por el investigador y de las decisiones teóricas que éste adopte (Hair, Anderson, Tatham & Black, 1999). Además, la nominación de los factores es un paso que, si bien se apoya en los contenidos aludidos por la formulación de los ítems, ofrece para muchos autores algunas dudas en virtud de la alta subjetividad implicada. Salvando estas cuestiones, es importante comprender que, teniendo en cuenta las limitaciones señaladas –presentes, por otra parte, en toda actividad científica– la atenta lectura de los resultados derivados del análisis factorial y del juicio experto –validez de contenido; ver apartado 2.2– permitirán que el usuario idóneo tome decisiones responsables sobre el uso o no uso de alguna técnica en particular en cada caso puntual.

### Validez convergente y discriminante

Establecidas ya las caracterizaciones básicas de estos aspectos teóricos de la validez, nos hallamos en condiciones de avanzar un paso más. Dentro de la validez de constructo es posible distinguir dos grandes aspectos: las evidencias de validez convergente y las evidencias de validez discriminante (García Cueto, 1993).

Cuando nos referíamos a la validez de constructo, aclarábamos que ella intentaba responder al interrogante de si la técnica de evaluación producida es el resultado de una adecuada operacionalización del constructo teórico que intenta medir; en términos sencillos, si resulta coherente con la teoría en que se basa. Ahora bien, dentro de cada diferente modelo teórico que ha construido la Psicología, existe una gran variedad de conceptos que se relacionan tanto por afinidad o contigüidad cuanto por diferencia u oposición. Así, las teorías producen cadenas conceptuales mediante las que los conceptos o constructos se definen, se caracterizan, se vinculan y se diferencian. Estas relaciones nos permiten entender cada concepto y discriminarlo de otros. Es por ello que en todo proceso de validación de constructo de un test interesa conocer si se han podido recoger evidencias de validez convergente y discriminante, es decir, evidencias de que el constructo medido por el instrumento *converge* en el mismo sentido que otra evidencia relacionada por similitud y, a la vez, que aparece evidencia discriminante, que se distingue teóricamente del concepto medido. A su vez, debe tenerse presente que todo estudio de validez de constructo tienen por objetivo recoger evidencias de validez teórica sobre un instrumento, pero, en última instancia, una multiplicidad de investigaciones de este tipo contribuyen, recíprocamente, a aportar evidencias acerca de la validez que la teoría misma aporta al explicar un fenómeno dado, es decir, cierta parte de la realidad. Avancemos un poco más.



Sin ahondar en sutilezas o profundizaciones que son irrelevantes para el alumno que recién se introduce en el tema, las *evidencias de validez convergente* son aquellas que se recogen cuando los resultados de un estudio de validez de constructo convergen en un mismo sentido, verificando la relación entre constructos vinculados teóricamente. Por ejemplo, hay evidencias de validez convergente cuando en un análisis factorial –recordemos que es uno de los métodos empleados para establecer evidencias de validez de constructo para los resultados arrojados por una prueba, aunque también podemos obtener tales resultados sin pasar por un procedimiento factorial - obtenemos un coeficiente de correlación bastante elevado– estimemos, por ejemplo, 0,90 - entre dos dimensiones o constructos relacionados, tales como ansiedad e insomnio, o depresión y anhedonia –incapacidad para sentir placer por diversas actividades–. Este índice de correlación cercano a 1 indicará una elevada covariación o asociación de ambas dimensiones teóricas, señalando su vinculación; si tal relación ha sido prevista por el modelo, entonces, estas evidencias –recogidas a partir de protocolos que contienen las respuestas de personas reales y concretas– reforzarán las afirmaciones postuladas por la teoría y, por ende, su capacidad explicativa sobre una porción de la realidad. Hablamos de validez convergente ya que ambas variables o dimensiones medidas –depresión y anhedonia; ansiedad e insomnio– se hallan asociadas en las respuestas brindadas por los sujetos, tal como la teoría ha postulado (García Cueto, 1993).

En el caso de las evidencias de *validez discriminante*, existirán tales resultados cuando obtengamos coeficientes de correlación relativamente bajos entre dimensiones o constructos diferentes del que se desea medir (Hogan, 2004), que el modelo ha concebido como relacionados teóricamente pero con una frecuencia de aparición conjunta en la realidad muy escasa (Cohen & Swerdlik, 2001); por ejemplo, depresión y manía en el mismo momento temporal –aunque amabas están presentes en los cuadros bipolares, ambas fases no se superponen en un mismo corte temporal dado; además, la teoría, si bien las diferencia, también las relaciona como alteraciones del estado de ánimo–.

Tal como comentábamos antes, las dos aristas del concepto validez de constructo detalladas hasta aquí –validez convergente y discriminante–, si bien se refieren primariamente a la técnica psicométrica que se intenta validar, contribuyen, en última instancia, a la validación de algunos aspectos del modelo teórico que la sustenta. Es justamente por referirse primariamente a la técnica, que algunos autores proponen la inclusión de las evidencias convergentes y disriminantes en el grupo de los aspectos de la validez empírica o de criterio (Hogan, 2004). Pero quedémonos, en este texto introductorio, con los consensos más tradicionales, entendiendo estas evidencias convergentes y discriminantes como aportes a las evidencias de validez de constructo.

### Otros estudios posibles

Otros estudios posibles que aportan evidencias a favor de la validez de constructo de un instrumento psicométrico son, por ejemplo, los meta-análsis, los estudios de evidencias de cambio pretest-postest y las matrices multi-método/multi-rasgo.

Los *meta-análisis* son investigaciones en las que se analizan diversos aspectos de estudios realizados por otros investigadores sobre un tema dado, en este caso, sobre un instrumento en particular. Los aspectos que se toman en cuenta y se discuten son, por ejemplo, de tipo metodológico, psicométrico, cultural, entre otros a ser considerados. Ello implica que toda la evidencia aportada –o, al menos, la más importante– sobre la validez teórica de un test determinado es minuciosamente revisada y puesta bajo una lupa muy crítica, con el fin de rescatar los resultados verdaderamente positivos y de prevenirse acerca de los negativos. La multiplicidad de estudios que se revisan hace que las conclusiones arrojadas por los desarrollos meta-analíticos sean de gran peso, pero existen ciertas limitaciones a ser tenidas en cuenta, como por ejemplo, la heterogeneidad de métodos y poblaciones con los que se haya trabajado, cuestión que podría estar dificultando la comparación de resultados, y el acceso real y efectivo que pueda tenerse a *todas* las publicaciones sobre el instrumento que se analiza (Glass, 1976).

Por su parte, los *estudios de evidencias de cambio pretest-postest* simplemente suponen que si las puntuaciones obtenidas en una prueba por una muestra de sujetos experimenta cambios por efecto de la aparición de cierto evento, entrenamiento o experiencia entre la administración del test y su nueva aplicación, estas variaciones –si son significativas– pueden convertirse en evidencias de validez de constructo (Cohen & Swerdlik, 2001). En términos simplificados, un test de comprensión lectora podría mostrar una mejora en sus resultados luego de un entrenamiento sistemático en esa destreza, una psicoterapia entre el pretest y el postest podría producir cambios en las puntuaciones en una escala de ansiedad. Claro está que el diseño de este tipo de investigaciones involucra una complejidad importante, ya que debería incluir a un grupo control que no reciba el entrenamiento –o la psicoterapia, según el ejemplo que se tome en cuenta– y ese grupo control no debería exhibir variaciones significativas en sus puntuaciones para que se verifique que el cambio ha sido dado por efecto de la intervención entre ambas evaluaciones.

Por último, las matrices multi-método/multi-rasgo implican un trabajo altamente complejo puesto que exigen la medición del mismo constructo en una muestra de individuos por medio de, al menos, dos vías, seleccionándose, asimismo, otros constructos diferentes pero relacionados con el que resulta de interés. Entonces, a la misma muestra de sujetos se les administran todos los tests destinados a medir el constructo central y todos los otros relacionados. De esta manera, al calcularse las correlaciones entre todas estas medidas, se obtiene evidencia empírica que apunta a las relaciones del constructo con otros, así como a su potencialidad de ser medido por medio de diversos métodos; por todo ello, este procedimiento de denomina multi-método (varios tests que miden el constructo que interesa y los otros relacionados) /multi-rasgo (todos esos constructos que ya nombramos: el de interés y los que con él se vinculan). Si puede probarse que un concepto teórico, medido por variedad de métodos guarda correlación con otros conceptos, tal como un modelo teórico dado prevé, pues entonces, verificar estas relaciones en situaciones reales con personas reales, aporta evidencias al aspecto teórico de la validez, es decir que se analiza la adecuación de los instrumentos psicométricos como medidas de un constructo dado que, a la vez, puede probarse que se halla relacionado con otros, teórica y empíricamente (Campbell & Fiske, 1959; Martínez Arias, 1995).



## 2.5 Aspectos de la validez vinculados con las características formales de la prueba

El último tipo de validez que examinaremos aquí es la *validez de facies o validez aparente*. *Facies* significa *rostro* o *cara* en latín (Bibliograf, 1977), y a ello, justamente, nos estamos refiriendo en este caso. La validez aparente tiene que ver con que el instrumento resulte válido a los ojos del examinado (Anastasi & Urbina, 1998; García Cueto, 1993), ya que si los materiales, la consigna, el estilo de respuesta o las condiciones propuestas para el examen dejan de aparecer como "serias" o adecuadas a su edad o sus características, podría suceder que los sujetos vieran afectada su actitud de respuesta, produciéndose un efecto contrario a lo deseado en aras del logro de un buen *rapport* (ver Cap. 1).

Hace algunas décadas era frecuente que muchos tests originalmente pensados para niños se extendieran a grupos de edades superiores, tales como los adolescentes, o aún, en algunos casos, a los adultos. El hecho de que en algunos de ellos no se adaptaran los materiales a los nuevos grupos a los que se destinaba el instrumento, hacía que se trabajara con estímulos –figuras, historias o contenidos– infantiles, decisión que frecuentemente generaba una actitud negativa en los adolescentes o adultos que respondían, puesto que entendían que se los estaba *tratando como a niños*, y ello influía en que no produjeran su mejor rendimiento. Algo similar sucede cuando parte del contenido de una técnica se publica en alguna revista de interés general, como entretenimiento para los lectores, contraviniendo las normas vigentes en nuestro país y en el resto del mundo. Como alguna vez ha sucedido, alguna persona lo ha respondido al leer la revista y luego se ha encontrado con idénticos materiales al concurrir a una evaluación psicológica en un gabinete, consultorio o en cualquier otro ámbito donde se realicen estas actividades; en estos casos, está claro que la imagen del profesional interviniente en tal proceso de evaluación, así como la de la herramienta empleada se habrán visto sumamente opacados ante tales circunstancias.

Es por ello que este aspecto es vigilado también, indirectamente, en el proceso de elaboración y de adaptación de las escalas psicométricas (ver Cap 5), en varios momentos tales como la redacción de los ítems, el juicio experto y la administración piloto. De esta manera, los materiales más adecuados a las características de los individuos a los que se los destina, logran que se trabaje de la mejor manera posible y por ende, obteniéndose la mejor actitud de respuesta que facilitará los mejores resultados.

### A modo de síntesis de lo hasta aquí desarrollado

Resumiendo, entonces, podemos comparar los aspectos de la validez relacionados con el contenido y con el modelo teórico (Fig. 1) así como los estudios concurrentes y predictivos correspondientes al concepto de validez empírica y a la validez de facies (Fig. 2).

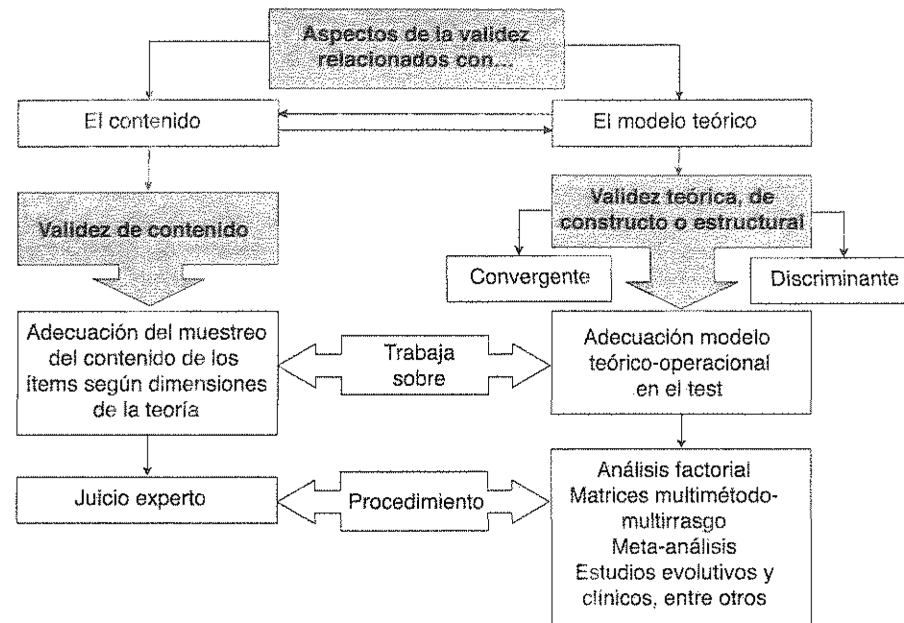



Fig. 1. Aspectos de la validez vinculados con el contenido de los ítems y con el modelo teórico.

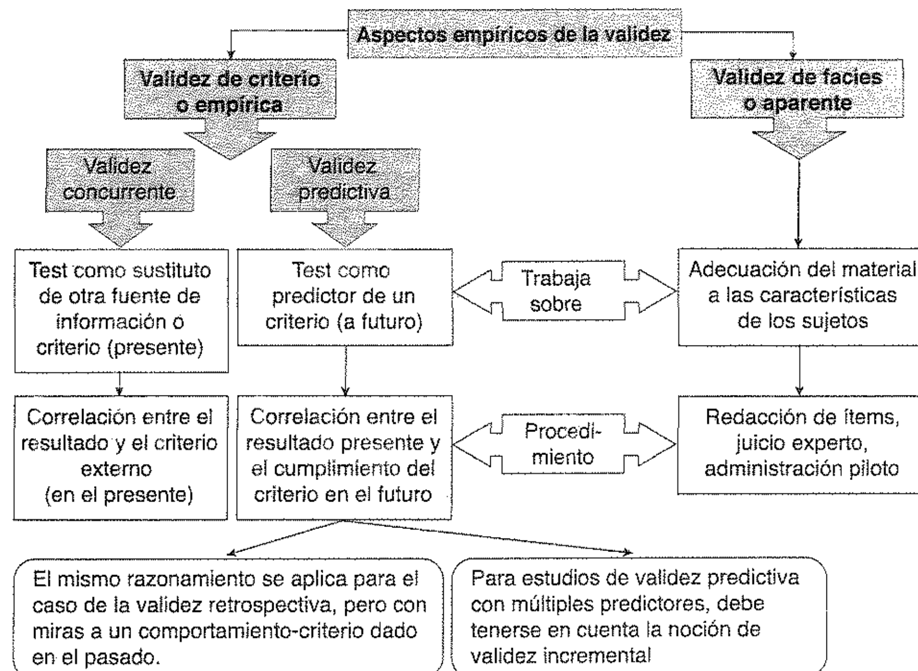



Fig. 2. Aspectos empíricos y formales del concepto de validez



Es importante destacar que un usuario de tests que conserve una postura responsable en su trabajo debería buscar en el manual del instrumento a emplear la mayor cantidad de estudios de validez posibles, si bien también resulta fundamental comprender que no necesariamente todos ellos habrán sido efectivamente determinados. Por ejemplo, tal vez una escala no tiene el propósito primario de predecir comportamientos, por lo que es posible que no tenga desarrollados estudios de validez predictiva. También es factible que no se efectúen investigaciones sobre la validez discriminante o sobre la validez concurrente por grupos contrastados, porque no sea de interés para los autores o porque resulte difícil llevarlas a cabo. Sin embargo, al menos deberíamos encontrar resultados que aporten evidencias de validez de constructo, de contenido y de validez concurrente, que asegure el uso práctico de la técnica en la situación presente.

Por otra parte, si bien es cierto que la validez de facies se vigila en varios momentos de la construcción del test, tal como hemos comentado, es poco frecuente que se la mencione en los manuales en forma explícita, ya que se infiere este trabajo cuando se menciona el proceso de redacción de ítems, de juicio experto y de pilotaje.

## 2.6 Sesgo y error sistemático

Para ir finalizando este capítulo, introduciremos brevemente al estudiante a la noción de *sesgo* (Anastasi & Urbina; 1998; Cohen & Swerdlik, 2001; García Cueto, 1993; Hogan, 2004; Martínez Arias, 1995), que es un concepto relacionado con el de validez; si bien el tema del sesgo se tratará en detalle en el Cap 5, es preciso señalar aquí su vinculación con la validez de los resultados aportados por un test.

El sesgo se define como un error constante o sistemático como opuesto al aleatorio o azaroso (ver Cap. 4), que impide la medición precisa e imparcial del constructo a evaluarse. Por acción de este sesgo, la probabilidad de éxito –o de un tipo de respuesta en especial si no se trata de una prueba de rendimiento– no es independiente del subgrupo poblacional al que pertenece el examinado. Ello significa que, aunque el instrumento haya sido estandarizado para un grupo poblacional dado, un subgrupo dentro de ese grupo mayor generará –probablemente– respuestas atípicas o no exitosas –si es un inventario de personalidad, por caso, o si se trata de una escala de desempeño, respectivamente– por acción de la pertenencia a ese subgrupo –conectada con la generación de un error sistemático– y no por otras razones.

Ese error sistemático –dado por cualquier componente inherente a la prueba, como por ejemplo, los materiales, el sistema de respuesta, los contenidos, las consignas u otros, se acumula– se da siempre de la misma manera y en el mismo sentido toda vez que el instrumento se administra a algún integrante de ese subgrupo. Por ello y a partir de ello, el test mide distintos constructos o genera distintos pronósticos, funcionando en forma diferencial en grupos disímiles, por razones ajenas a la variable que la escala está destinada a medir.

No abundaremos en diferenciaciones entre los dos tipos de sesgo que la psicometría –de intersección y de pendiente– prevé puesto que no consideramos que ello sea relevante para nuestros estudiantes, pero –refiriendo al lector al Cap. 5 en el que se profundiza este tema–, simplemente dejaremos sentada la relación del concepto de sesgo –y de error sistemático– con el concepto de validez. Un test que funciona en forma diferencial en un subgrupo de población dado deja de arrojar resultados válidos para los miembros de ese subgrupo, ya que estas personas obtienen puntuaciones

distintas en la medición de la variable por factores ajenos a ella, es decir que sus desempeños no obedecen a diferencias reales en el nivel de la variable a medirse sino a cuestiones vinculadas con su inserción en un subgrupo. Así, por ejemplo, si una escala de inteligencia infantil se compone mayoritariamente de materiales relacionados con situaciones en las que se utilizan computadoras personales y videojuegos, aquellos niños menos expuestos o directamente no expuestos a tales estímulos rendirán por debajo del resto de sus pares, pero no ya por tener una inteligencia menor sino por efectos de hallarse menos familiarizados con los materiales empleados en los ítems del instrumento. Por lo tanto, la validez de los resultados obtenidos se verá afectada, y este error será sistemático ya que cada vez que un niño con menor acceso a PCs y videojuegos sea evaluado, su rendimiento será más bajo a causa de su menor nivel de entrenamiento en tales materiales y no por efecto de la variable a ser medida por la técnica, que es la inteligencia.

Debe recordarse, además, que el concepto de sesgo –relacionado con la validez– se vincula al de error sistemático, en tanto que el de confiabilidad y el de error de medición se relacionan con el de error aleatorio (ver Cap. 4).

## Resumen general y comentarios finales

Más allá de las distinciones antes mencionadas, vale la pena tener en cuenta que el concepto de validez ha sufrido variaciones a lo largo de la historia de la Psicometría. Si bien la definición más clásica estipula que validez es el **grado en que un instrumento mide aquel constructo que pretende medir** (Garrett, 1937), la dificultad radica en operacionalizar ese grado de relación, y allí, precisamente, es donde se ha dado la importante evolución histórica a la que aludimos (Martínez Arias, 1995).

Inicialmente, primaba una perspectiva utilitaria, en tanto se buscaba referir la validez a los propósitos prácticos, de uso o empleo del test, en relación a su poder predictivo con respecto a ciertos comportamientos que sus resultados eran capaces de pronosticar o predecir (Bingham, 1937; Cureton, 1950; Guilford, 1946). Validez era sinónimo de correlación entre las puntuaciones derivadas de una prueba psicométrica y alguna otra medida del constructo que esa prueba intentaba medir (Angoff, 1988; Martínez Arias, 1995). A partir de aquí, en los años ´50s los estándares internacionales comenzaron a distinguir la validez concurrente de la predictiva, diferenciando la correlación del test con un criterio en la situación presente de su potencialidad para la predicción de un comportamiento futuro (APA, 1954; AERA, APA & NCME, 1974).

Con el correr de los años, se empezó a mirar la parte teórica del asunto, volcándose la focalización a la adecuación test-modelo teórico, y no solamente vigilando el aspecto práctico o empírico inherente al uso del instrumento; así, validez comenzó a definirse como el grado en que el contenido de la escala representa una muestra satisfactoria de todos los contenidos posibles que pueden ser incluidos en la evaluación del constructo; de esta forma, surgió el concepto de validez de contenido que ya hemos examinado en los apartados anteriores.

El concepto de validez aparente apareció de la mano de Mosier (1947), aunque debe reconocerse que nunca se le ha otorgado demasiada relevancia a lo largo de la historia, en tanto que los cuatro tipos principales de los que hablamos hoy en día –de contenido, de constructo, concurrente y predictiva– se sistematizaron a partir de 1954, en los estándares APA ya mencionados. A partir de este nuevo ordenamiento de la nomenclatura, en el que el término validez de constructo realizó su aparición, comenzó a perfilarse la idea de que la validez no puede ser expresada por medio de un índice único, por lo que requiere de una multiplicidad de estudios dedicados a establecer todos sus tipos.



Pocos años más tarde y ya en relación con este concepto de validez de constructo que iba cobrando cada vez más importancia, Campbell y Fiske (1959) diferenciaron la validez convergente y la discriminante al interior de la validez de constructo –aunque recordemos que autores como Hogan (2004) las ubican en el grupo de validez de criterio–. Luego de muchos años de debate alrededor de estas etiquetas clásicas y de otras no tan difundidas, se ha llegado a establecer una nomenclatura de uso más corriente, que es la que aquí se ha presentado.

Baste, simplemente, con decir que si bien es necesario determinar todos los tipos de evidencias de validez que resulte posible examinar, la validez de constructo se constituye en el concepto que unifica e integra todas las cuestiones de contenido y empíricas o de criterio (validez concurrente y predictiva) en un cuadro holístico que permitirá la puesta a prueba de hipótesis racionales sobre relaciones y distinciones teóricamente relevantes en cuanto al constructo que pretende ser medido (Messick, 1980). Es decir que la validez teórica se vuelve el concepto integrador que subsume a los otros tipos, aunque cada uno vale por sí mismo y debe ser analizado por separado, pero sin perder de vista que, en última instancia, un test es una colección de indicadores que resultan de la operacionalización de un constructo teórico dado (ver Cap. 1).

Como resumen general de este capítulo, puede decirse que (Fig. 3):

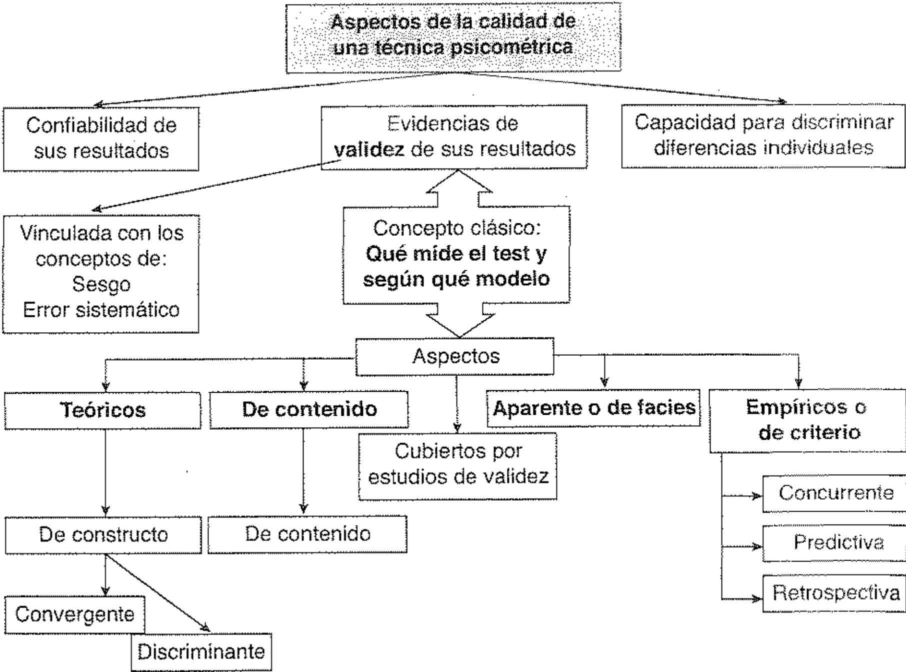



Fig. 3. Aspectos del concepto de validez y estudios correspondientes

Debe destacarse que los tipos o aspectos de validez a los que hasta aquí se ha hecho referencia no agotan todos los clasificados en las obras de psicometría, puesto que existen nomenclaturas diferenciales según los autores que se consulten. En el presente texto sólo nos hemos referido a aquellos más difundidos en la visión más clásica de la psicometría, que bastan para introducir al estudiante en los conceptos básicos que un usuario de tests debe manejar. Sin embargo, debe tenerse en cuenta que otros autores aluden, por caso, a la validez muestral y curricular, que se cuentan como parte de la validez de contenido y la validez factorial, que se incluye en la validez de constructo (García Cueto, 1993). Para este autor, la *validez muestral* alude a dos aspectos diferentes relacionados con el contenido del instrumento: por un lado, a la relevancia de los contenidos previstos en los ítems y, por otro, al examen de en qué medida el contenido de los reactivos cubre todos los aspectos del campo o dominio teórico que se quiere evaluar. En este sentido, esta definición se está refiriendo a las dimensiones de la variable contempladas en el contenido de los reactivos, por lo que podemos apreciar que se solapa con la definición misma de *validez de contenido*.

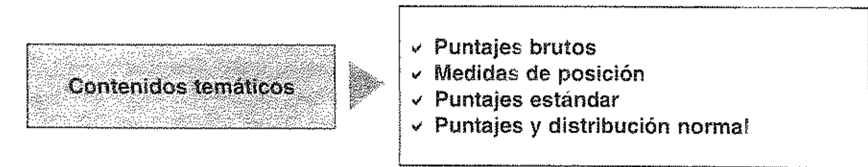
La *validez curricular*, por su parte, se relaciona con los contenidos incluidos en diversas pruebas de admisión que se elaboran para decidir el ingreso de postulantes a distintas instituciones, por lo general, educativas. Se relaciona, así, con la pertinencia de esos contenidos en cuanto a la currícula vigente en dichas organizaciones.

Por último, la validez factorial parece coincidir con la validez de constructo, estudiada desde el punto de vista del análisis factorial, tal como hemos visto en el apartado *2.4* de este capítulo. Sin embargo, los términos validez de constructo y validez factorial no son sinónimos: la expresión *validez factorial*, que alude al aspecto teórico de la validez (validez de constructo), está cayendo en desuso, ya que sólo se restringe a examinar la adecuación modelo-operacionalización en el test mediante una metodología factorial, dejando de lado los otros métodos disponibles (matrices multimétodo-multirrasgo, meta-análisis, grupos contrastados, estudios evolutivos, etc.). Más bien, en este caso, actualmente se tiende a hablar de *estudios factoriales que aportan evidencias de validez de constructo.*

*Capítulo 3*

# *Las puntuaciones de los test*

Marcelo Antonio Pérez




## 3.1 Los puntajes brutos

Luego del amplio panorama abierto en el primer capítulo, éste pondrá el foco en un tema central de la psicometría, los puntajes. Como se irá desarrollando, los hay de distinto tipo y se analizarán los que se consideran básicos.

Ya se ha hecho referencia a los cuatro niveles de medición conceptualizados por Stevens en 1946, quien a su vez definió la medición como *la asignación de números o símbolos a objetos o fenómenos siguiendo ciertas reglas*. De acuerdo con esta definición, los ***instrumentos de medición psicológica,*** o sea, las pruebas psicométricas, quedaron caracterizadas como aquellas que permiten relevar y/o procesar información psicológica en números. Pero conviene aquí distinguir entre las palabras número y numeral.

### Numerales y niveles de medición

Los ***numerales*** son símbolos numéricos (1, 2, 3, I, II, III, etc.) mientras que el ***número*** es la cantidad que estos símbolos representan. Si los signos numéricos se asignan a las distintas modalidades que puede tener una variable no cuantitativa, estos símbolos solo funcionan como numerales - no como números- y en este caso no tiene sentido hacer operaciones matemáticas entre ellos. Solo si los numerales guardan una relación fija cuantitativa entre ellos permiten realizar las operaciones matemáticas básicas y se llamarán números.

Es decir, los numerales pueden representar cualidades o cantidades pudiéndose distinguir:

1) Numerales nominales: en este caso solo sirven para *nombrar* las distintas modalidades de la variable, y estaríamos haciendo una medición cualitativa (clasificación). Este uso es el que da lugar al nivel de medición ***nominal***.
2) Numerales ordinales: aquí podemos *posicionar* las modalidades de la variable. A esta medición se la denomina semi-cuantitativa. En este caso el nivel de medición es el ***ordinal***.
3) Numerales cardinales: aquí se *cuantifican* las modalidades de la variable, el numeral es un número. La medición es cuantitativa. De acuerdo a este nivel cuantitativo es que quedan definidos los niveles de medición de ***razones*** e ***intervalar*** ya desarrollados en el capítulo anterior que solo difieren en la naturaleza de su cero, por ello suele tratárselos indistintamente como un único nivel de medición denominado ***escalar***.
Las operaciones matemáticas que pueden hacerse entre numerales dependen de su tipo. En la fig. 1.5. (pág. 19) el lector podrá encontrar una síntesis de los distintos niveles de medición mientras que en la siguiente tabla se los encuentra relacionados con las operaciones admitidas en cada nivel de medición y el tipo numeral respectivo.

Tabla 3.1. Operaciones válidas según nivel de medición del numeral

| Tipo Numeral | Operaciones admitidas entre sus numerales | Nivel de medición |
|---|---|---|
| Nominales | = ≠ | Nominal |
| Ordinales | = ≠ > < | Ordinal |
| Cardinales (números) | = ≠ > < + - x | Intervalar |
| | = ≠ > < + - x ÷ | Razones o cocientes |

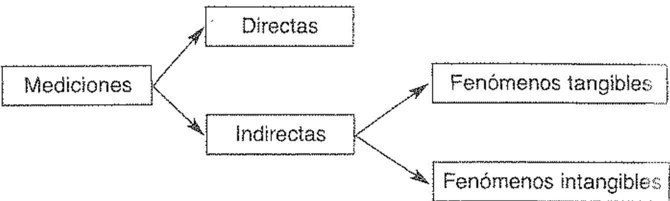
Las reglas a las que refiere Stevens en su definición apuntan a la existencia de un isomorfismo entre los números y los fenómenos psicológicos que se pretenden medir. El grado de isomorfismo que presenten definirá las propiedades matemáticas que podrán aplicarse entre esos números, ya que estas deben poder verificarse empíricamente entre los fenómenos psicológicos por ellos representados. Así, por ejemplo, solo se podrán sumar dos números asignados a distintas intensidades de un fenómeno si es que se verifica empíricamente que el fenómeno en cuestión es sumativo.

## Mediciones psicológicas

En la ciencia al realizar una medición debe especificarse no solo el valor medido sino también el error con que este se calcula. Hay varias modalidades para expresar dicho error, las cuales serán desarrolladas con detalle en el siguiente capítulo. Digamos aquí que dicho error tiene relación - entre otros determinantes- con la facilidad de acceso a la información sobre lo que quiere medirse. Desde este punto de vista, las mediciones pueden clasificarse en directas e indirectas.

Mediciones directas serán aquellas donde el fenómeno a medir pueda observarse a través de los sentidos, no hace falta hacer ninguna inferencia, se puede "contar", calcular, o bien se puede comparar lo que se desea medir con un objeto o fenómeno similar. El sexo de una persona, su edad, la cantidad de palabras que es capaz de






recordar, son ejemplos de mediciones directas. Nótese que en este tipo de medidas el error puede o no existir, y en el primer caso suele ser sencillo calcularlo o estimarlo.

En las mediciones indirectas, en cambio, el fenómeno a medir no puede evaluarse en forma directa sino que se hace necesario el uso de un instrumento para "materializarlo" y de este modo asignarle números. Dentro de los fenómenos psicológicos de medición indirecta conviene distinguir dos tipos:

- los fenómenos tangibles, como suelen ser las variables físicas o psicofísicas (por ejemplo el tiempo que tarda un sujeto en resolver un problema),
- los fenómenos intangibles, es decir que se infiere su misma existencia de los indicadores que se le atribuyen en su operacionalización.

En el caso de las mediciones indirectas de fenómenos tangibles al error que se podría cometer con una medición directa se le debe agregar el error que introduce el instrumento. Por ejemplo, si se mide el tiempo que tarda un sujeto en resolver un problema deberá agregarse el error del reloj o cronómetro. Todo instrumento que se agregue a una medición es una posible fuente de error que deberá contemplarse.

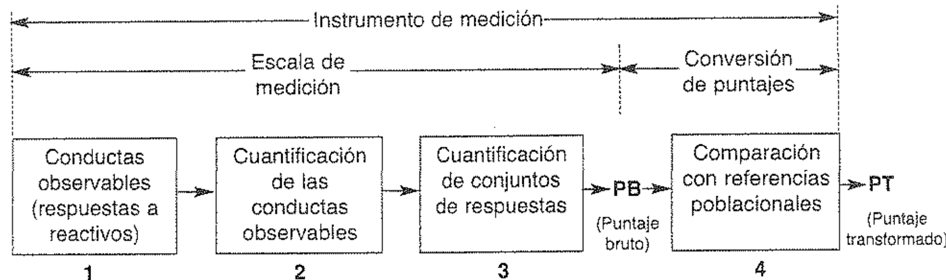
Cuando las mediciones son indirectas y de fenómenos intangibles el error se maximiza. En el capítulo 1 se mencionó que a estos fenómenos psicológicos no observables en sí mismos, que tienen existencia ideal, teórica, se los denomina ***constructos***, y que éstos para ser medidos deben ser operacionalizados, es decir, traducidos en un conjunto de comportamientos a cuya cantidad, frecuencia o intensidad se les pueda asignar un número o numeral. En este caso al error cometido por el instrumento en esta asignación, deberá agregársele el producido en su operacionalización y validación. En el próximo capítulo se profundizarán aspectos vinculados al error. En el próximo capítulo se profundizarán aspectos vinculados al error.

En este capítulo en particular se hará referencia a los instrumentos ***cuantitativos o semi-cuantitativos;*** es por eso que interesa en este punto destacar cuatro aspectos que se deben tener en cuenta en su proceso constructivo:

1. Cómo se operacionaliza el constructo: relevar cuales son las conductas o cogniciones observables que dan cuenta del mismo (indicadores) y lo distinguen de otros.
2. Cómo se fomenta que las conductas a evaluar se manifiesten, y cómo se les asignan números que representen ya sea su intensidad o frecuencia o, al menos, su presencia-ausencia.
3. Cómo operar con esos números para lograr, en lo posible, uno (o pocos) que sean la expresión cuantitativa y/o cualitativa más acabada posible del constructo que se pretende medir.
4. Cómo transformar los números obtenidos en resultados que sean significativos para su interpretación.

En el siguiente gráfico se presenta un diagrama de bloques en un instrumento de medición indirecta genérico; cada uno de estos bloques representa uno de los cuatro aspectos anteriormente mencionados, ya formalizados como componentes del mismo.

Gráfico 3.1. Cuatro aspectos que se deben tener en cuenta en la construcción de un instrumento de medición




Los aspectos referidos tanto a generar reactivos y como cuantificarlos –expresados por los dos primeros bloques del gráfico– ya fueron desarrollados extensamente en el primer capítulo y se desarrollan también en el capítulo 5. En éste se desarrollarán aspectos ligados a los dos últimos bloques, es decir, a cómo se cuantifican los conjuntos de respuestas obteniendo lo que se denomina *puntaje bruto*, y cómo se lo valora a través de los denominados puntajes *transformados* y los *baremos*.

## El puntaje bruto

Ya se ha dicho que los fenómenos intangibles, objetos de mediciones indirectas, se denominan constructos, y que la operacionalización de los mismos permite encontrar sus indicadores.

Por otro lado, cualquier constructo operacionalizado u otro fenómeno que pueda ser medido con más de una modalidad, es decir, que tiene más de un estado posible, en el contexto de la investigación y medición psicológica, se denomina ***variable.*** Así pues, referiremos como ***variable psicológica*** a los fenómenos de interés de la psicología que tienen variabilidad, que se manifiestan con más de una modalidad, sean estas cualitativas o cuantitativas. En el ámbito científico las variables psicológicas se las nomina y expresa con una letra minúscula (generalmente x, y, z).

Si se pueden contar o ponderar las conductas observables atribuidas a una variable psicológica (ver como ejemplo Ansiedad en el capítulo 1 pág. 35), el resultado es una "medida del fenómeno". Este principio es la base funcional de la mayoría de los instrumentos de medición psicológica indirectos.

Se ha mencionado también en el capítulo 1, que una prueba psicométrica incluye la noción de escalamiento, definido entonces *como la posibilidad de convertir o traducir las respuesta brindadas por los sujetos a una puntuación.* Ampliando el concepto, Amón (1978) define como escala de medición psicológica al conjunto de modalidades



(estados distintos) de un constructo psicológico vinculados unívocamente a un conjunto de números (numerales) distintos. Se podría también expresar que es un conjunto de números asignados unívocamente a una variable psicológica, y para ser unívocos los numerales asignados deberán cumplir con dos propiedades: ser ***exclusivos*** (que cada una de las modalidades solo pueda ser representada por un numeral) y ***exhaustivos*** (que todas las modalidades posibles tengan asignado un numeral).

Una vez asignados los numerales a los ítems según lo visto en el capítulo anterior, es deseable operar sobre ellos de forma tal de obtener otros nuevos que tengan una significación más robusta e isomorfa con la variable que se quiere evaluar. Así, si se determinó que un conjunto de respuestas están evaluando la misma variable, podríamos por ejemplo sumar los números que les asignamos a cada una de ellas a fin de obtener un nuevo valor que represente el total de dicho conjunto. Debe tenerse en cuenta que para poder sumar o hacer otras operaciones matemáticas entre los numerales asignados a las respuestas, será preciso que cumplan los siguientes requisitos:

a) cuantifiquen (o al menos semi-cuantifiquen) la misma variable.
b) lo estén haciendo siguiendo las mismas reglas (por ejemplo en cualquier respuesta, un número mayor indique mayor presencia de la variable).
c) que los números asignados representen valoraciones isomorfas de la variable (por ejemplo, items más difíciles deben tener asignados números mayores, síntomas más prototípicos o representativos de la variable, mayor ponderación que los síntomas accesorios, etc.).

Si se logran estos requisitos, entonces, este nuevo número representará de una forma más acabada la cantidad/ cualidad/ frecuencia de la variable en cuestión. La demostración o el consenso necesario para asignar números a las respuestas de los reactivos y operar con ellos, forma parte del proceso de estandarización y validación de la técnica.

A este nuevo número, que sintetiza y representa la cantidad/cualidad/frecuencia de la variable y que ha sido el resultado de este proceso de medición, se lo suele denominar ***puntaje bruto, crudo o directo. El puntaje bruto es, entonces, un número que representa una cuantificación de la variable o constructo a medir.*** Gran parte de los instrumentos de medición tiene como resultado al menos un puntaje bruto, aunque pueden no tenerlo.

El puntaje bruto es, en general, el resultado final de la ***escala de medición.***

He aquí un ejemplo sencillo de una escala de medición: se desea medir el constructo/variable "memoria inmediata verbal" en un sujeto; para tal fin se le lee al mismo un listado de 30 palabras, solicitándole luego que verbalice todas las que recuerde. Para que la cantidad de palabras que recuerda el examinado sea indicador de su "cantidad de memoria inmediata verbal" hubo primero que operacionalizar el constructo y validar este supuesto. Si esto fue así, y se asigna un punto (1) a cada palabra bien recordada y cero (0) a las que no recordó o recordó mal, la suma de estos resultados será un nuevo número que variará de 0 a 30: cuanto más elevado sea este valor –nominado puntaje bruto– más "cantidad" de la variable a medir indicará. Se obtuvo así una escala de medición, –que, en este caso, será una colección de 31 números que representan unívocamente distintas modalidades ("cantidades") del constructo (0 a 30)–.

## Nivel de medición del puntaje bruto

El nivel de medición que puede ser utilizado en los numerales obtenidos por la combinación de varios ítems guarda relación con el nivel de medición que poseían los ítems a combinar y con la forma de hacerlo.

Gran parte de los instrumentos de medición psicológica obtiene su puntaje bruto como resultado de la simple suma de números asignados a las respuestas de los reactivos que miden la misma variable, o de su conteo. No obstante conviene aclarar que esta no es la única alternativa, pudiéndose utilizar operaciones matemáticas más complejas (restas, multiplicaciones, cocientes entre otras).

Como ya se indicó, si los numerales asignados a las respuestas corresponden a un nivel nominal, entonces no son números y por lo tanto no se puede realizar operaciones matemáticas sobre ellos; en tal caso, puede contarse la cantidad de respuestas en un sentido u otro, y el resultado de este conteo sí lo será. En cambio si los números son tales, se puede operar sobre ellos con las operaciones lícitas al nivel de medición que poseen.

El valor resultante de una combinación de ítems suele mejorar el nivel de medición que tiene el ítem aislado. Por ejemplo, si los ítems tienen nivel ordinal, el resultado obtenido al combinarlos adecuadamente es ordinal o incluso podría ser intervalar. Si los ítems son intervalares, la combinación resultante seguramente también lo será, mientras que si los ítems indican presencia ausencia (por ejemplo nivel nominal de una dicotómica) se puede contar la cantidad de frases que respondió en un sentido (u otro) y así obtener una escala cuantitativa. En conclusión, el número resultante de agrupar ítems suele tener, al menos, el nivel de medición de los ítems con que se compuso y con frecuencia, lo supera.

Es muy deseable que la escala obtenida tenga el máximo nivel de medición posible, ya que esto permite trabajar los números en niveles matemáticos y estadísticos más elaborados y precisos. Por ello es que los diseñadores de instrumentos no ahorran esfuerzos al momento de diseñar las escalas para que los resultados alcancen niveles de medición más elevados.

A tal fin, en muchas ocasiones, escalas que estrictamente corresponden al nivel ordinal pero que tienen alta probabilidad de que los intervalos sean iguales, son tratadas como intervalares. Esto se da con frecuencia en aquellas escalas que tienen rangos de resultados amplios, ya sea porque están conformadas por muchos ítems, porque tienen muchas opciones de respuestas posibles, o por ambas cosas. Si bien son consideradas a efectos prácticos como intervalares, no hay que olvidar que en sentido estricto, debería demostrarse que los intervalos son iguales para poder incluirla dentro de este nivel de medición.

## Valoración del puntaje bruto

El puntaje bruto –número que ha resultado de la escala de medición– suele ser poco claro para la evaluación por parte del usuario de la técnica, ya que si bien cuantifica o cualifica el constructo, por sí mismo no ofrece suficiente información con respecto a la magnitud de la medida obtenida. Para comprender si el puntaje bruto es alto, bajo o intermedio se requiere de un sistema de referencia externo, generalmente una comparación con los valores que comúnmente obtienen los demás sujetos.



Así por ejemplo, si una persona obtuviera 15 puntos en el instrumento para medir memoria al que ya nos referimos, este resultado puede dar una primera impresión de que el sujeto ha obtenido un valor de memoria "término medio", ya que el valor de 15 está en la mitad de la escala. Pero si en promedio las demás personas obtienen 5 puntos, estos 15 indicarían que el sujeto en cuestión se desempeñó mucho mejor que los demás, dando la impresión de que tiene "mucha memoria"; en cambio si el valor medio que obtienen las otras personas es de 25 puntos, estos 15 indicarían lo contrario.

La dificultad principal para la comprensión del significado de este número radica en que en psicología no existen unidades de medición como en las ciencias exactas. Si hubiera una unidad de medición de la memoria (como el metro en la longitud, el gramo en el peso) sería claro entender si un puntaje de 15 es alto o no.

Lo que se estila hacer para valorar al puntaje bruto es compararlo con otros valores que permitan contextualizarlo, como se hizo en el ejemplo, con el promedio de palabras memorizadas. Los valores más utilizados para contextualizar los puntajes brutos son las frecuencias (absolutas, relativas, acumuladas, mediana), la media (o promedio) y el desvío estándar, obtenidos del conjunto de datos de una población.

Cuando a estos valores se los utiliza para convertir los puntajes brutos en otros, a los últimos se los conoce como *Puntajes Transformados*. Nótese que los nuevos números que se asignarán a los puntajes brutos ya no son producto de la cuantificación directa del constructo a medir, es decir no son el producto de una escala de medición, sino que se trata de nuevos números, fruto de la comparación de los puntajes brutos con referencias poblacionales. Estos nuevos números no expresan cuánto de la variable puntuó el examinado, sino cuánto puntuó en relación a los demás.

Esta reconversión de un puntaje bruto a uno transformado ya no es una "medición" en el sentido de la asignación de números al fenómeno que se quiere medir; ahora se trata de una conversión del número que realmente cuantifica al constructo (puntaje bruto) en otro más útil para interpretar (puntaje transformado). Por este motivo es que es poco práctico y hasta confuso hablar de niveles de medición en los puntajes transformados ya que estos son compuestos.

Es oportuno remarcar que los instrumentos de medición suelen estar, entonces, conformados por una escala de medición y un dispositivo de conversión para la valoración, cuyos resultados dan puntajes relacionados con los primeros, pero de distinta naturaleza. Por ello es que el gráfico 2.1, puede ilustrarse de una manera resumida en este nuevo gráfico.

Grafico 3.2. Distintos puntajes según el proceso de construcción




A su vez, los puntajes transformados pueden clasificarse en dos grandes tipos: las Medidas de Posición y los Puntajes Estándar. Los siguientes apartados serán dedicados a estos tipos de puntuaciones.

## 3.2 Medidas de posición

Este tema implica el repaso de algunos conceptos estadísticos y para exponerlos más claramente se utilizará como ejemplo un instrumento genérico.

Para comenzar digamos que lo ideal para establecer comparaciones y valorar el puntaje obtenido por un sujeto es hacerlo con los puntajes de la ***población***, entendiendo esta última como el conjunto de todos los sujetos con los que se desea comparar al primero. A los valores estadísticos que se obtienen de esos puntajes (como son el promedio, la mediana, el desvío estándar, entre otros) se los denomima ***parámetros***; es decir, lo ideal sería comparar el puntaje del sujeto a examinar con estos parámetros. Así en la prueba de memoria ejemplificada anteriormente, para saber qué magnitud tiene un determinado puntaje se debería comparar el mismo, por ejemplo, con el promedio obtenido de los puntajes de todos los sujetos que compartan características similares a quién lo obtuvo: igual edad, sexo, condición socioeconómica, educación, etc.

En la gran mayoría de los casos es muy poco práctico –en muchos otros, imposible– lograr obtener los puntajes de toda la población para realizar dicha comparación, sea porque ésta es muy numerosa o inaccesible o porque es muy caro el proceso; por ello es necesario recurrir a subconjuntos de la misma llamados ***muestras***.

### Muestreo

Al calcular los estadísticos de estas muestras, utilizando recursos de la rama de la estadística llamada estadística inferencial, se pueden estimar los parámetros. Lo más importante para que estos parámetros estén adecuadamente estimados, es haberlos calculado con muestras representativas de la población.

Este último punto es tan central para reducir el error, que una rama de la estadística tiene como objetivo el estudio de distintas técnicas de muestreo y el cálculo de los errores que pueden cometerse en las inferencias que de ellas se obtienen.

El acceso a estas muestras de las cuales se obtienen los valores de referencia forma parte de las tareas del equipo de los psicometristas que diseñan los instrumentos, es decir, los instrumentos editados ya vienen con sus valores de referencia incluidos, o con tablas u otros dispositivos para obtener los puntajes transformados. Como el destinatario de este libro es el evaluador –el que utiliza el instrumento en ámbitos de aplicación–, es que no se desarrollarán aquí las distintas técnicas de muestreo, que el lector interesado podrá consultar en textos especializados (Amón, 1980; Botella & San Martín,1997; Coolican,1997; Cortada de Kohan, 1994; Pagano, 2006).

No obstante, el usuario de una técnica debe tener conocimientos sobre la validez de los resultados que obtiene, y por ello conocer juiciosamente el ajuste de los valores obtenidos de las muestras que tienen los manuales frente a una determinada medición. Al menos, debe asegurarse que el sujeto a examinar sea similar a los que conformaron la muestra, es decir, que éste hubiera sido una persona elegible para la misma. Cuanto más parecido sea el examinado a los sujetos que conformaron la muestra, menos error de medición habrá, y se obtendrán resultados más ajustados desde un punto de vista métrico.

Si la muestra no es del todo adecuada y no se dispone de otra ni de otro instrumento, habrá que corregir los resultados usando todas las herramientas posibles: la



calidad de la medición habrá caído, y habrá un mayor trabajo de interpretación y evaluación para compensar tal caída.

Si la muestra es inadecuada, el instrumento de medición no podrá usarse como tal (los puntajes transformados no son ajustados y no deberán utilizarse), ya que aún siendo útil la escala de medición, no lo es la valoración. En este caso, la habilidad y experiencia del evaluador para la interpretación y valoración de los resultados son cruciales.

### Organización de los puntajes: frecuencias

Una vez que se ha seleccionado la muestra, se administra el instrumento en cuestión a todos sus integrantes –siguiendo la consigna al "pié de la letra" como se indicó en capítulo 1 pág. 26 –, y se calculan los resultados obtenidos por cada uno de los sujetos. Con esos puntajes, convenientemente tratados, se calculan los estadísticos y se estiman los parámetros. Estos valores son presentados en los manuales para uso del evaluador, en forma de tablas que permiten la conversión de los puntajes brutos a puntajes transformados. Dichas tablas reciben el nombre de ***baremos o normas estadísticas***.

Para ejemplificar este proceso, recurriremos nuevamente a la escala de memoria ya mencionada. Suponemos que el instrumento ha sido administrado a una muestra de 120 personas y por lo tanto se han obtenido 120 resultados, que se presentan en la siguiente Tabla.

Tabla 3.2. Resultados de una muestra de 120 sujetos

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 8 | 12 | 10 | 13 | 15 | 17 | 14 | 9 | 13 | 10 | 14 | 19 | 12 | 10 | 16 | 13 | 14 | 18 | 12 |
| 11 | 21 | 3 | 13 | 16 | 6 | 15 | 23 | 12 | 16 | 15 | 8 | 13 | 17 | 18 | 11 | 12 | 13 | 11 | 9 |
| 14 | 10 | 20 | 9 | 11 | 14 | 17 | 13 | 14 | 18 | 11 | 20 | 12 | 14 | 13 | 15 | 27 | 10 | 14 | 16 |
| 13 | 11 | 12 | 14 | 13 | 15 | 16 | 6 | 10 | 22 | 12 | 15 | 9 | 15 | 16 | 15 | 11 | 17 | 14 | 13 |
| 16 | 14 | 13 | 15 | 16 | 14 | 12 | 13 | 17 | 14 | 11 | 16 | 13 | 14 | 6 | 11 | 19 | 12 | 2 | 15 |
| 12 | 15 | 8 | 14 | 11 | 10 | 15 | 12 | 8 | 13 | 17 | 15 | 12 | 14 | 15 | 9 | 16 | 18 | 14 | 10 |

Puede observase que los resultados presentados en la Tabla 2. son difíciles de interpretar, por lo que es razonable buscar alguna forma de ordenarlos. Uno de los criterios más sencillos para agruparlos, es contar cuantas personas obtuvieron cero puntos, cuantas un punto, cuantas dos, y así sucesivamente hasta 30 puntos, que es la puntuación máxima en esta escala. De este modo, los resultados pueden ser sintetizados como lo muestra la Tabla 3.3

Tabla 3.3. Agrupación de los resultados por frecuencias

| Puntaje | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cantidad de casos | **0** | **0** | **1** | **1** | **0** | **0** | **3** | **0** | **4** | **5** | **8** | **10** | **13** | **15** | **18** | **14** |

| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10** | **6** | **4** | 2 | **2** | **1** | **1** | 1 | **0** | 0 | **0** | 1 | 0 | 0 | **0** |

Si se quitan de la tabla las columnas donde al puntaje bruto que lo encabeza tuvo ausencia de personas que lo hayan obtenido (cantidad= 0), se obtiene una tabla más ordenada y pequeña:

Tabla 3.4. Agrupación de resultados por frecuencias distintas de cero

| Puntaje | 2 | 3 | 6 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 27 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cantidad | **1** | **1** | **3** | **4** | **5** | **8** | **10** | **13** | **15** | **18** | **14** | **10** | **6** | **4** | **2** | **2** | **1** | **1** | **1** | **1** |

A las cantidades de casos correspondiente a cada uno de los distintos puntajes se las denomina ***frecuencia*** (también ***frecuencia absoluta***) de dicho puntaje. Así por ejemplo, la frecuencia o frecuencia absoluta con que se obtuvo 9 puntos en esta muestra fue igual a 5, es decir hay cinco personas que obtuvieron 9 puntos. Estas frecuencias (número de casos) se expresan con la letra **n**, y cada puntaje (la variable psicológica) se lo expresa con la letra **x.** Así entonces para x= 9 hay una frecuencia n=5.

A la **cantidad total** de datos que conforman la muestra se denomina con la letra **N**. En este ejemplo, N= 120.

Se denomina ***frecuencia relativa*** a la frecuencia absoluta dividida la cantidad total de datos que conforma la muestra, y se expresa con la letra **p**. La frecuencia relativa expresa entonces que proporción hay de determinado puntaje respecto del total de puntajes o datos que conforman la muestra.

$$p = \frac{n}{N} \text{ en el ejemplo } p = \frac{n}{N} = \frac{5}{120} = 0{,}041$$

Para mayor claridad, a la frecuencia relativa se la multiplica por 100, obteniendo entonces **la frecuencia relativa porcentual.**

$$p\% = \frac{n}{N} \times 100 = 0{,}041 \times 100 = 4{,}1\%$$

Es decir que el 4,1 % de las 120 personas de la muestra, obtuvieron 9 puntos en la escala de memoria.

Por otro lado, es posible demostrar que, si se suman las frecuencias relativas de todos los puntajes posibles, el resultado será 1; análogamente, si se suman todas las frecuencias relativas porcentuales de todos los puntaje posibles, el resultado será 100%.

Nótese que la frecuencia arroja información sobre qué tan recurrente o común es ese puntaje. Así el puntaje x=14 lo han obtenido n=18 personas, siendo su p%= 15%, lo cual significa que ésta puntuación se da con mayor frecuencia que el puntaje x= 9 al que le corresponde un p%= 4,1%.

## Distribución de frecuencias: mediana

Una manera de mejorar el ordenamiento de los datos cuando se trabaja con una variable de nivel ordinal, intervalar o de cocientes, es utilizar las frecuencias acumuladas. Para ello se ordenan los resultados de menor a mayor y se calcula la frecuencia obtenida por cada uno de ellos. Se llamará ***frecuencia acumulada*** a la frecuencia que tiene un determinado puntaje más (sumadas a ella) las frecuencias que tienen todos los resultados menores a él, es decir, se calculan la cantidad de datos que se hallaron con el valor en cuestión o con valores inferiores.

Así en el ejemplo, la frecuencia acumulada para el puntaje 6 es la frecuencia que tiene ese puntaje más las frecuencias de los puntajes 5, 4, 3, 2, 1 y 0 . La frecuencia acumulada se puede calcular para las frecuencias absolutas (**na**), relativas (**pa**) y relativas porcentuales (**pa%**). La tabla siguiente indica las frecuencias acumuladas con los datos del ejemplo que se va llevando.



Tabla 3.5. frecuencias absolutas, relativas acumuladas y porcentuales

| X | n | p | p% | na | pa | pa% |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0,0083 | 0,83 | 1 | 0,0083 | 0,83 |
| 3 | 1 | 0,0083 | 0,83 | 2 | 0,016 | 1,6 |
| 4 | 0 | 0 | 0 | 2 | 0,016 | 1,6 |
| 5 | 0 | 0 | 0 | 2 | 0,016 | 1,6 |
| 6 | 3 | 0,025 | 2,5 | 5 | 0,033 | 3,3 |
| 7 | 0 | 0 | 0 | 5 | 0,033 | 3,3 |
| 8 | 4 | 0,033 | 3,3 | 9 | 0,075 | 7,5 |
| 9 | 5 | 0,041 | 4,1 | 14 | 0,116 | 11,6 |
| 10 | 8 | 0.066 | 6,6 | 22 | 0,183 | 18,3 |
| 11 | 10 | 0,083 | 8,3 | 32 | 0,266 | 26,6 |
| 12 | 13 | 0,108 | 10,8 | 45 | 0,375 | 37,5 |
| 13 | 15 | 0,125 | 12,5 | 60 | 0,5 | 50 |
| 14 | 18 | 0,15 | 15 | 78 | 0,65 | 65 |
| 15 | 14 | 0,116 | 11,6 | 92 | 0,76 | 76 |
| 16 | 10 | 0,083 | 8.3 | 102 | 0.85 | 85 |
| 17 | 6 | 0.05 | 5 | 108 | 0,9 | 90 |
| 18 | 4 | 0,033 | 3,3 | 112 | 0,933 | 93,3 |
| 19 | 2 | 0,016 | 1,6 | 114 | 0.95 | 95 |
| 20 | 2 | 0.016 | 1,6 | 116 | 0,966 | 96,6 |
| 21 | 1 | 0,0083 | 0,83 | 117 | 0,975 | 97,5 |
| 22 | 1 | 0,0083 | 0,83 | 118 | 0,983 | 98,3 |
| 23 | 1 | 0,0083 | 0,83 | 119 | 0,991 | 99,1 |
| 24 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 25 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 26 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 27 | 1 | 0,0083 | 0,83 | 120 | 1 | 100 |
| 28 | 0 | 0 | 0 | 120 | 1 | 100 |
| 29 | 0 | 0 | 0 | 120 | 1 | 100 |
| 30 | 0 | 0 | 0 | 120 | 1 | 100 |

En la última columna se encuentran las frecuencias acumuladas porcentuales, que indican que porcentaje de puntajes brutos es igual o menor al que corresponde a dicha frecuencia acumulada porcentual. Así, por ejemplo para el puntaje bruto Pb= 8 corresponde una pa%= 7,5 %, es decir que el 7, 5 % de los sujetos de la muestra obtuvo 8 puntos o menos, que es lo mismo que decir que el 7,5% de los puntajes de esa muestra es de 8 puntos o menos.

Puede verse que para el Pb: 13 corresponde una frecuencia acumulada de 50%, es decir que la mitad de los puntajes (o de los sujetos de la muestra) obtuvo 13 puntos o menos. A ese valor que deja por debajo la mitad de las puntuaciones –que es lo mismo que decir la mitad de los sujetos de la muestra–se denomina ***Mediana***. La mediana en el ejemplo será M*e*= 13

Todos los puntajes transformados llamados *Medidas de Posición* se basan en las frecuencias acumuladas porcentuales. Los más destacados en psicología son el Percentil, el Decil y el Cuartil; en general a este tipo de medidas se las denomina ***cuantiles*** o ***fractiles***.

## Percentil

Como se ha visto, la mediana es el valor que divide al conjunto de los datos en dos mitades con la misma cantidad de datos (en este caso puntajes brutos), y en el ejemplo se corresponde al valor Pb= 13. Análogamente se puede calcular el puntaje que corresponde a la mediana de cada una de estas dos mitades en que se dividieron los datos con la mediana, obteniéndose dos nuevos puntajes que dividen cada una de las dos mitades en otras dos mitades, con la misma cantidad de datos cada una. Estos dos valores agregados al de la mediana dividen en cuatro partes los datos originales, con un 25% de puntajes o casos en cada una de ellas; nótese que tendríamos tres puntajes que dividen los datos en cuatro agrupamientos, cada una de ellos con el 25 % de los datos.

Si se deseara dividir la distribución en 100 partes en cada una de las cuales se encuentre la misma cantidad de casos, hacen faltan 99 valores. Esos puntajes que dividen la distribución en 100 partes con el 1% de los casos en cada una de ellas se denominan percentiles. El ***percentil*** –también llamado centil por muchos autores– expresa qué porcentaje de mediciones de la muestra tiene por debajo o en el mismo valor cada puntaje bruto.

Además de la denominación, centil o percentil, en la bibliografía también varía la forma de abreviarlo: a veces se le pone el signo "%" detrás del número, otras lo abrevian como %il; en ocasiones suele aparecer también con la letra p (ó C) y con subíndice el percentil ( por ejemplo percentil 50 como $p_{50}$.) y también con la palabra percentil delante del número. Todas estas variantes se producen en castellano, debido a que el término original proviene del inglés y en las lenguas sajonas se utilizan números ordinales para expresarlos, cosa muy difícil e inusual en castellano. Así por ejemplo el percentil 25º se expresa –twenty fifth centile– (percentil vigésimo quinto).

El percentil es muy usual porque tiene numerosas ventajas. La primera de ellas es ser un valor fácil y claro de interpretar, ya que su número indica el porcentaje de sujetos de la muestra que obtuvieron el mismo valor o menor en la variable medida.

Por ejemplo, si se observa la tabla 3.6, al percentil 50 –que es lo mismo que decir la mediana– le corresponde el Pb:13, y al percentil 90 el Pb: 17. Es decir que una persona que obtuvo 13 puntos ha sacado un puntaje tal que deja por debajo de él al menos



al 49% de los puntajes de la muestra, y quién saca 17 se ha ubicado un puntaje que lo ubica por encima de al menos el 89 % . En términos prácticos, se dice que supera al 50 ó 90 por ciento respectivamente.

Otro aspecto ventajoso de los percentiles es que no resulta complicado hacer un listado de puntajes brutos, asignarle a cada uno el percentil correspondiente y presentarlo como una tabla, que como se dijo en el capítulo 1 pag. 48 se la denomina ***Baremo***. La misma Tabla 3.6 es el Baremo que muestra las correspondencias entre los puntajes brutos y los percentiles del ejemplo que se viene desarrollando.

Tabla 3.5. (repetición)

| X | n | p | p% | na | pa | pa% |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0,0083 | 0,83 | 1 | 0,0083 | 0,83 |
| 3 | 1 | 0,0083 | 0,83 | 2 | 0,016 | 1,6 |
| 4 | 0 | 0 | 0 | 2 | 0,016 | 1,6 |
| 5 | 0 | 0 | 0 | 2 | 0,016 | 1,6 |
| 6 | 3 | 0,025 | 2,5 | 5 | 0,033 | 3,3 |
| 7 | 0 | 0 | 0 | 5 | 0,033 | 3,3 |
| 8 | 4 | 0,033 | 3,3 | 9 | 0,075 | 7,5 |
| 9 | 5 | 0,041 | 4,1 | 14 | 0,116 | 11,6 |
| 10 | 8 | 0,066 | 6,6 | 22 | 0,183 | 18,3 |
| 11 | 10 | 0,083 | 8,3 | 32 | 0,266 | 26,6 |
| 12 | 13 | 0,108 | 10,8 | 45 | 0,375 | 37,5 |
| 13 | 15 | 0,125 | 12,5 | 60 | 0,5 | 50 |
| 14 | 18 | 0,15 | 15 | 78 | 0,65 | 65 |
| 15 | 14 | 0,116 | 11,6 | 92 | 0,76 | 76 |
| 16 | 10 | 0.083 | 8,3 | 102 | 0,85 | 85 |
| 17 | 6 | 0,05 | 5 | 108 | 0,9 | 90 |
| 18 | 4 | 0,033 | 3,3 | 112 | 0,933 | 93,3 |
| 19 | 2 | 0,016 | 1,6 | 114 | 0,95 | 95 |
| 20 | 2 | 0,016 | 1,6 | 116 | 0,966 | 96,6 |
| 21 | 1 | 0,0083 | 0.83 | 117 | 0,975 | 97,5 |
| 22 | 1 | 0,0083 | 0,83 | 118 | 0,983 | 98,3 |
| 23 | 1 | 0,0083 | 0,83 | 119 | 0,991 | 99,1 |
| 24 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 25 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 26 | 0 | 0 | 0 | 119 | 9,991 | 99,1 |
| 27 | 1 | 0,0083 | 0,83 | 120 | 1 | 100 |
| 28 | 0 | 0 | 0 | 120 | 1 | 100 |
| 29 | 0 | 0 | 0 | 120 | 1 | 100 |
| 30 | 0 | 0 | 0 | 120 | 1 | 100 |

Tabla 3.6. Baremo

| PB | Percentil |
|---|---|
| 0 -2 | 1 |
| 3-5 | 2 |
| 6-7 | 4 |
| 8 | 8 |
| 9 | 12 |
| 10 | 19 |
| 11 | 27 |
| 12 | 38 |
| 13 | 50 |
| 14 | 65 |
| 15 | 76 |
| 16 | 85 |
| 17 | 90 |
| 18 | 93 |
| 19 | 95 |
| 20 | 96 |
| 21 | 97 |
| 22 | 98 |
| 23-30 | 99 |

El baremo con puntajes brutos y percentiles se puede confeccionar a partir de una tabla de puntajes brutos y frecuencias acumuladas porcentuales.

La interpretación de los percentiles de estos baremos es fácilmente comprensible. Retomando el ejemplo, las personas que obtuvieron 16 puntos habrían "superado en memoria" al 85 % de la muestra (como vimos, en términos estrictos, al 84 %).

Nótese que a los puntajes brutos que van desde el cero al dos le corresponde el mismo percentil, es decir que se encuentran dentro del mismo rango percentilar, de la misma manera que lo están los valores 23 a 30. Es importante destacarlo ya que indica que si un sujeto obtuvo una puntuación de 23 o más puntos, el percentil que obtendrá como resultado no varía, es decir que para los valores extremos de esta escala de medición, el percentil no refleja las variaciones de puntajes. Esto es una limitación de las medidas de posición en general. Esta característica, sumada al hecho de no asignar un valor en forma unívoca a cada estado medido, hacen que este tipo de medidas pierdan –en casos como éste– el carácter de escala (se recuerda que la escala de medición tiene una relación *unívoca* entre un puntaje y una modalidad).

Otra limitación del uso de los percentiles en los instrumentos de medición psicológica es que difícilmente los baremos incluyen los 99 valores; puede verse en el ejemplo que se viene desarrollando que solo se lograron 18 percentiles para los 31 puntajes brutos, lo que implica que al pasar los puntajes brutos a percentiles se produzca una pérdida del rango de amplitud del instrumento, lo cual puede llevar a errores de lectura al utilizarlos.

Nótese, además, que los percentiles, en realidad, son 99 valores de frecuencia acumulada que marcan 100 posiciones. No obstante, es usual que se usen decimales y así se exprese por ejemplo percentil 99,5 o 99,9, aunque en términos estrictos estos no son percentiles (no dividen la distribución de frecuencia en 100 partes iguales) sino frecuencias acumuladas. La gran difusión de la utilización de los percentiles con decimales ha popularizado su uso y también a veces su redondeo, y así por ejemplo al "percentil" 99,99 se lo ha llamado "percentil 100", y a valores tan bajos como "percentil 0,01" se los ha dado a llamar "percentil 0", aunque en términos estrictos no existan tales valores de percentiles, sino solo de frecuencias acumuladas. Aún aceptando esta distorsión del concepto percentil originario, se debe recordar que en todo caso jamás un percentil puede ser mayor a 100 ni menor a cero, límites reales de las frecuencias acumuladas.

A continuación se vuelve a presentar la tabla baremo correspondiente al ejemplo, con datos iguales a la anterior, pero en ésta se han cambiado las columnas por las filas para mayor claridad.

Tabla 3.7. Baremo con columnas y filas intercambiadas

| PB | 0-2 | 3-5 | 6-7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23-30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentil | 1 | 2 | 4 | 8 | 12 | 19 | 27 | 38 | 50 | 65 | 76 | 85 | 90 | 93 | 95 | 96 | 97 | 98 | 99 |

Es muy frecuente que las tablas de este tipo se simplifiquen aún más, ya sea porque se han logrado aún menos de los 19 valores percentilares de este ejemplo, o por motivos de practicidad. Así, no debería extrañar al usuario que, por ejemplo, la tabla terminara siendo publicada en los manuales de la siguiente manera.



Tabla 3.8. Baremo simplificado

| PB | 8 | 9 | 11 | 13 | 15 | 17 | 19 |
|---|---|---|---|---|---|---|---|
| Percentil | 5 | 10 | 25 | 50 | 75 | 90 | 95 |

Obtenido un puntaje bruto del instrumento, para transformarlo con esta tabla se debe proceder teniendo en cuenta las siguientes viñetas.

- Si el sujeto obtuvo alguno de los puntajes brutos que figuran en la fila *Puntaje bruto (PB)* su percentil es el que figura inmediatamente abajo en la fila *Percentil*. Por ejemplo, si el examinado obtuvo un puntaje bruto de 17 puntos, se puede observar que el percentil que le corresponde es de 90; puede leerse entonces que el examinado *"supera" en "memoria" al 90 % de la muestra con la que se construyó el baremo.*
- El valor inferior de puntajes brutos que posee la tabla es de 8 puntos, al que le corresponde un percentil de 5. En caso de que algún sujeto obtenga cualquier puntaje bruto inferior a este valor, se deberá interpretar que el examinado obtuvo un *percentil inferior a 5: su puntaje en memoria está por debajo del 5% de la muestra.*
- El valor superior de los puntajes brutos de la tabla es de 19 puntos. En caso de que se obtenga un puntaje bruto superior a ese valor, se deberá interpretar que el percentil en el que se ubica el examinado es *superior al percentil 95, es decir supera en memoria a más del 95 % de la muestra.*
- Si la persona hubiera obtenido un puntaje bruto intermedio entre dos de los valores expresados en la tabla, como por ejemplo el valor 12 (estaría entre el PB 11 al que le corresponde el percentil 25 y el PB 13 al que le corresponde el percentil 50) se deberá interpretar el resultado como *entre el percentil 25 y 50, es decir supera en memoria del 25 al 50 % de la muestra.* No se debe interpolar un valor percentilar entre dos de los valores que figuran en la tabla, salvo que se explicite algún procedimiento para hacerlo en el instrumento utilizado: la interpolación supone que hay un incremento lineal entre los dos valores de interpolación lo cual primero debería demostrarse, además de arrojar muchas veces un número con decimales que genera una impresión ficticia respecto al grado de exactitud que tiene la técnica: por otro lado, los percentiles son números enteros. Si por algún motivo se debiera decidir entre los dos valores percentilares, se utilizará el más cercano a la media (en este caso el percentil 50).

Es también frecuente, y sobre todo en escalas con rangos pequeños de resultados, que dos o más puntajes brutos estén asignados en el baremo al mismo percentil. Supóngase que se hace un nuevo baremo de la misma técnica que se viene ejemplificando, pero ahora en una muestra de individuos con severas dificultades de memoria. En dicha muestra, dada la dificultad de recordar de los participantes, todos los valores descenderán, pudiendo llegar a convertirse el baremo anterior en otro similar al que se muestra en la siguiente Tabla.

Tabla 3.9. Baremo con otra muestra

| PB | 2 | 3 | 3 | 4 | 5 | 6 | 6 |
|---|---|---|---|---|---|---|---|
| Percentil | 5 | 10 | 25 | 50 | 75 | 90 | 95 |

En este caso, se observa que al Pb= 3 le corresponde un percentil de 10 y también el percentil 25: se deberá leer, entonces, que su percentil está entre 10 y 25. Análogamente, a un Pb= 6 le corresponde el percentil 90 o más. Si se debiera elegir un solo percentil representativo de dicho puntaje, siguiendo la regla antes descrita se asignará el percentil 25 para el PB= 3, y el percentil 90 para el PB= 6.

Otro aspecto de los baremos de percentiles es que suelen utilizarse intervalos para interpretarlos, a los que se les asignan categorías (bajo, alto, término medio, etc.). Los intervalos y las categorías más usuales son los siguientes.

Tabla 3.10. Rangos usuales en la interpretación de los percentiles

| Percentil entre | Valor... |
|---|---|
| 1-10 | **Muy bajo** |
| 11-25 | **Bajo** |
| 26-75 | **Término medio** |
| 76-90 | **Alto** |
| 91-99 | **Muy alto** |

Finalmente, si bien el percentil presenta la ventaja de ser sencillo para interpretarlo, se deberá tomar nota de algunas características cuyo desconocimiento lleva a errores en su lectura. Las siguientes son las más destacables.

- La gran mayoría de las técnicas de evaluación psicológica está muy lejos de ofrecer en sus baremos realmente 100 posiciones. Este es el principal motivo para no hacer interpolaciones y para usar los intervalos de interpretación antedichos.
- Los percentiles extremos suelen ser mucho más imprecisos que los centrales, vale decir, cuando el percentil en que se encuentra el examinado es muy bajo o muy alto. Sensibles variaciones del constructo (expresadas en amplias variaciones de los puntajes brutos) pueden verse reflejadas en pequeñas o inexistentes fluctuaciones del percentil que les corresponde.
- Los extremos percentilares no representan el mínimo ni el máximo de la variable que el instrumento puede evaluar, sino los mínimos y máximos de la comparación con la muestra.
- El percentil no es un porcentaje, sino una medida de posición, por lo cual nunca es mayor a 99 (ó 100 en caso de la frecuencia acumulada).
- El rango percentilar indica el porcentaje de sujetos de la muestra que ha sido superado por el número del percentil, pero no necesariamente que el complemento a 100 de ese número "lo supera". Así por ejemplo un sujeto que sacó un percentil de 90 ha superado en su puntuación al 90 % de la muestra, pero no se pueda afirmar que es superado por el 10 %; solo se puede indicar que está entre el 10 % de los sujetos que han superado al 90 % de la muestra.



Decil y cuartil

Los deciles son puntajes análogos a los percentiles, pero en lugar de tener un rango de 99 posiciones, tienen uno de 9. Estos nueve valores son aquellos que dividen a los datos en 10 conjuntos de igual cantidad – de forma análoga a como lo hacía la mediana en dos conjuntos-, y se obtienen también de las frecuencias acumuladas. Hay pocas escalas de medición en la actualidad que los utilicen, siendo la más destacada en nuestro medio el test 16 PF –hoy 16 PF5– con la que se evalúan los factores de personalidad propuestos por Cattell (Cattell 1975; Russell & Karol, 2000).

Las características de los deciles son las siguientes.

Valor mínimo: 1
Valor máximo: 9
Valor medio: 5

Estos puntajes comparten las mismas ventajas (y desventajas) que los percentiles, siendo aún más fácil calcularlos. Al tener solo 9 valores, su utilización se restringe a medidas más gruesas que en las que se utilizan los percentiles. Es por ello que en las escalas que se desea mayor precisión se usan los percentiles, aunque debe quedar claro que en muchas escalas disponibles se han usado percentiles dada su gran difusión, aunque estrictamente hubiera sido mucho más adecuado a su nivel de precisión utilizar deciles o, incluso, cuartiles, que se mencionan más adelante.

Multiplicando los deciles por 10 se obtendrá su rango percentilar, y éste se corresponde con los percentiles múltiplos de 10 respectivos (10, 20,...90). Si por ejemplo, al decil 2 le corresponde un rango percentilar de 20, quien obtiene un puntaje decil 2 "supera" en la variable medida al 20% de la muestra con la que se lo está comparando.

Por último, análogamente a lo sucedido con los percentiles, en algunos instrumentos aparecen los deciles con decimales.

Los cuartiles son otra medida de posición pero menos usual en evaluación psicológica, ya que sirven para dar resultados muy "gruesos". Dividen los resultados solo en cuatro agrupamientos de igual cantidad de datos, basándose en los percentiles 25, 50 y 75, a los que se los llama primer, segundo, tercer y cuarto cuartil. En general son medidas que son más útiles para trabajar con datos estadísticos o tomar decisiones sobre muestras, que en la construcción de instrumentos de evaluación.

Finalmente, se aclara que existen más puntajes que tienen concepciones similares a las mencionadas, como las estaninas (o estanueves) entre otros, pero la poca difusión actual de los mismos en nuestro medio hace inconveniente su tratamiento en este capítulo.

## 3.3 Puntajes Estándar

Los puntajes estándar se obtienen mediante un cálculo matemático por el que se logra comparar el puntaje bruto evaluado en un sujeto con el valor medio y el desvío estándar previamente calculados en una muestra. Para entender con profundidad por qué se utilizan estos dos estadísticos (media y desvío estándar), será necesario repasar algunos conceptos de estadística.

Clásicamente para explicar el concepto de puntaje estándar se parte de la distribución de frecuencias llamada normal conocida también como curva o campana de Gauss, ya que el uso más frecuente de estos hace referencia a esta distribución. No obstante, esto puede llevar a la noción de que los puntajes estándar están necesariamente ligados a esa distribución de frecuencias, afirmación que no es sostenible de hecho. Por estos motivos, y para evitar que el lector asocie los puntajes que se desarrollarán en esta sección directamente con la distribución normal de frecuencias, se dejará para más adelante este tópico (apartado 3.4.), donde se completarán los conceptos que se comienzan a desarrollar aquí.

Un aspecto importante a destacar de los puntajes estándar es que como se obtienen de una fórmula matemática aplicada a los puntajes brutos, a cada uno de estos le corresponde un único puntaje estándar. Es decir, los puntajes transformados– a diferencia de las medidas de posición– mantienen el aspecto unívoco que caracteriza a las escalas de medición.

### Puntaje diferencial: uso de la media

Al referirnos a los puntajes brutos fue aclarado que estos arrojan muy poca información sobre su magnitud, cosa que dificulta su valoración al evaluador, pero también fue dicho que esta mejora si cada puntaje bruto obtenido es comparado con el valor promedio de una población, permitiendo valorar si ese puntaje es alto o bajo o, al menos, si es mayor o menor que el promedio.

Para obtener el puntaje medio, llamado ***media o promedio*** es necesario, primero, administrar el instrumento a una muestra de sujetos y obtener los puntajes de cada uno de los individuos que la componen. Hecho esto, la media se obtiene sumando todos los puntajes y dividiendo el resultado de esa suma por la cantidad de sujetos evaluados. Estadísticamente se la expresa como $\overline{X}$ y su fórmula de cálculo es la que figura en el siguiente recuadro.

$$\overline{X} = \frac{\sum PB}{N}$$

*donde y $\sum$ es la sumatoria, o "suma de....”; PB son los Puntajes brutos, N la cantidad de sujetos de la muestra.*

Nota: dado que a las variables –sean psicológicas o no– se le suelen asignar letras minúsculas como x, y, z, el lector podrá encontrar en otras bibliografías la misma expresión con **x** en vez de PB (puntajes brutos).

Si al puntaje bruto obtenido por un sujeto se le resta el valor de la media, se obtiene un nuevo puntaje cuyo valor indica cuán apartado del valor promedio está el puntaje bruto en cuestión. Este nuevo puntaje que combina el puntaje bruto con la media se denomina ***puntaje diferencial***, y tiene la propiedad de que cuanto más grande es su valor –sea este negativo o positivo– mayor será la distancia del puntaje bruto respecto del valor promedio; tiene a su vez la propiedad de que cuando el puntaje bruto en cuestión vale cero éste coincide con la media.



$$Pd = PB - \overline{X}$$

*donde Pd es el Puntaje diferencial, PB es el Puntaje bruto y $\overline{X}$ es la media*

Si el puntaje bruto de un sujeto fuera superior al valor promedio, el puntaje diferencial que le corresponderá será un valor positivo. Por el contrario, si el puntaje diferencial es negativo, indica que el puntaje bruto es inferior al puntaje medio.

La ventaja del uso de los puntajes diferenciales con respecto a los puntajes brutos es que informan si la medida está por encima o por debajo de la media con solo ver su signo; también indican, con su valor absoluto, que tan lejos se está del valor promedio.

Sin embargo, la dificultad que tienen es que mantienen la limitación de los puntajes brutos respecto de la valoración de su magnitud. Si, por ejemplo, se tienen dos instrumentos A y B, que evalúan inteligencia, uno con un valor medio de 10 puntos y el otro de 100 puntos, el obtener en ambos un puntaje diferencial Pd= 8 tiene muy distinto significado: como el valor es positivo, se sabe que en ambos casos el puntaje es de 8 puntos por encima de la media, pero nada se puede decir sobre si esto es mucho o poco. Véase el siguiente recuadro.

| | | | | |
|---|---|---|---|---|
| *Pd= PB-$\overline{X}$* | A | Pd= 18-10= 8 | B | Pd= 108-100=8 |

En los instrumentos A y B se ha obtenido el mismo puntaje diferencial (8 puntos), pero en términos relativos puede observarse que en el instrumento A el valor obtenido (18 puntos) está “muy lejos” de la media (10 puntos) y en el caso B el valor obtenido (108 puntos) está “cerca” del valor de la media (100 puntos).

Una forma usual de poder observar la magnitud que representa ese puntaje diferencial consiste en referirla a (dividirla por) un valor conocido y estable, obteniendo de esta forma los ***puntajes diferenciales relativos***. Para mayor claridad en su lectura, a la proporción obtenida de esa división se la multiplica por 100, obteniéndose los ***puntajes diferenciales relativos porcentuales***.

El valor de referencia podría ser la misma media, u otros valores significativos tales como la mediana, la moda, el máximo. Si utilizamos la media como valor de referencia, los puntajes diferenciales relativos porcentuales de los ejemplos A y B serían los siguientes.

$$Pd\% = \frac{PB-\overline{X}}{\overline{X}} \text{ x } 100$$

$$\text{En } \mathbf{A} \text{ sería: } Pd\% = \frac{18-10}{10} \text{ x } 100 = 80\%$$

$$\text{y en } \mathbf{B} \; Pd\% = \frac{108-100}{100} \text{ x } 100 = 8\%$$

La variación de 8 puntos respecto de 10 representa una variación del 80%, en cambio, ese mismo valor (8) en 100 solo significa el 8 %. Evidentemente el mismo puntaje diferencial en un instrumento y en otro con distintas medias, no expresa efectos similares, es decir, no son comparables.

## Puntaje z: uso de media y desvío estándar

En el apartado anterior ha quedado claro que el puntaje diferencial ofrece más información que el puntaje bruto, más aún, al dividirlo por un valor de referencia. Sin embargo, calcularlo a partir de la media, como fue ejemplificado, no es lo óptimo ya que la media es un valor que no ofrece información sobre cuan cercanos o alejados a ella están los valores.

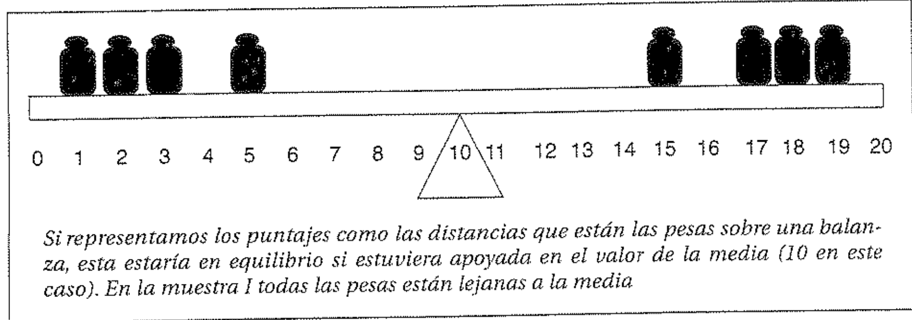
Volviendo al ejemplo, en el instrumento A, el promedio de 10 puede haberse obtenido a partir de una muestra de sujetos, en la que la mitad de ellos obtuvo puntajes muy "alejados" de la media (en el ejemplo puntajes alrededor de 5 y de 15 puntos) o bien en una muestra que arrojó valores muy "próximos" a la media (la mitad alrededor de 9 y la otra mitad alrededor de 11 puntos).

A los efectos de dejar en claro estos conceptos, se pondrá el foco en dos situaciones ejemplificadas con muestras de solo ocho casos, aunque esta cantidad de sujetos es muy reducida para realizar cualquier estudio psicométrico. En las siguientes tablas se presentan los datos de ambas muestras, la I y la II, y los gráficos respectivos que permiten analizarlas.

Tabla 3.11. Muestra I de 8 puntajes

| Muestra | $PB_1$ | $PB_2$ | $PB_3$ | $PB_4$ | $PB_5$ | $PB_6$ | $PB_7$ | $PB_8$ | $\Sigma$ | $\bar{X}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| I | 2 | 3 | 5 | 1 | 17 | 18 | 15 | 19 | **80** | **10** |

Gráfico 3.3.



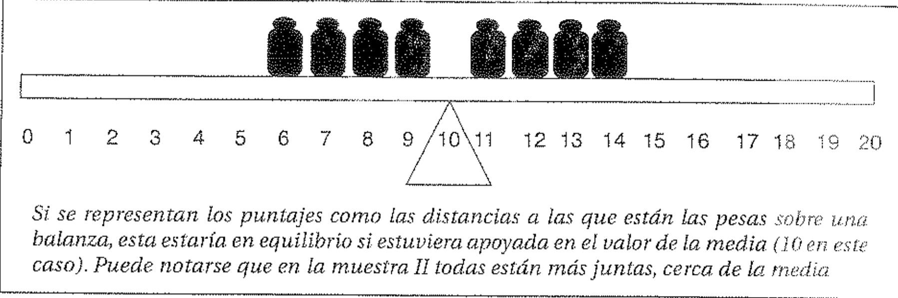

*Si representamos los puntajes como las distancias que están las pesas sobre una balanza, esta estaría en equilibrio si estuviera apoyada en el valor de la media (10 en este caso). En la muestra I todas las pesas están lejanas a la media*

Tabla 3.12. Muestra II de 8 puntajes

| Muestra | $PB_1$ | $PB_2$ | $PB_3$ | $PB_4$ | $PB_5$ | $PB_6$ | $PB_7$ | $PB_8$ | $\Sigma$ | $\bar{X}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| II | 6 | 9 | 7 | 8 | 12 | 11 | 13 | 14 | **80** | **10** |



Gráfico 3.4.




*Si se representan los puntajes como las distancias a las que están las pesas sobre una balanza, esta estaría en equilibrio si estuviera apoyada en el valor de la media (10 en este caso). Puede notarse que en la muestra II todas están más juntas, cerca de la media*

Obsérvese que el puntaje medio de 10 puntos se obtiene tanto con los datos de la muestra I como con los de la II.

En la muestra I, el puntaje bruto de 18 –que ya veníamos ejemplificando– es un valor cercano a los valores altos de esta muestra con la que se ha obtenido la media (los valores 17, 18, 15 y 19), lo que indicaría que es un puntaje frecuente de obtener.

En la muestra II, el mismo valor 18, está muy alejado de los valores con los que se compara (el más cercano es el 14). Es decir, es un valor mucho menos frecuente –atípico–, que ningún integrante de la muestra ha obtenido.

No obstante estas diferencias, en ambos casos el puntaje diferencial que se obtiene haciendo el correspondiente cálculo es *Pd*= 8 y el *Pd%*= 80%

De este modo se puede observar que si se calcula el puntaje diferencial y el diferencial relativo porcentual con la media en el cociente –como se ha hecho-, se obtiene en ambos casos el mismo puntaje diferencial relativo porcentual. Por lo tanto, podemos afirmar que este tipo de puntaje nada indica sobre si el valor con el que se obtuvo es común o si es atípico. He aquí una primera limitación del uso de la media en el cociente.

Por otra parte, la media es un valor que se ve sensiblemente afectado por la variación de los valores extremos (se dice que no es una medida robusta). Así por ejemplo, si en la muestra II se agrega un solo puntaje de valor 20, el nuevo cálculo de la media presentaría los datos siguientes.

Tabla 3.13. Muestra II más un caso extremo

| Muestra II + un caso extremo de 20 puntos ($PB_9$) | $PB_1$ | $PB_2$ | $PB_3$ | $PB_4$ | $PB_5$ | $PB_6$ | $PB_7$ | $PB_8$ | $PB_9$ | $\Sigma$ | X |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 9 | 7 | 8 | 12 | 11 | 13 | 14 | **20** | **100** | **11,1** |

El hecho de agregar un solo puntaje bruto igual a 20 ha hecho variar la media de 10 a 11,1. Esta inestabilidad es una segunda limitación del uso de la media como valor de referencia.

Para superar las limitaciones que tiene el uso de la media como cociente para dividir los puntajes diferenciales, se usa, entonces, otro estadístico que toma en cuenta la dispersión que tienen los valores respecto de la media.

Para obtenerlo se parte de los puntajes diferenciales que indican la distancia que tiene cada puntaje bruto respecto de la media, y un promedio de ellos sería una forma razonable de obtener una medida de cuán dispersos o alejados están estos valores. En

el ejemplo de las balanzas equivaldría a hacer el promedio de las distancias de las pesas al punto de equilibrio (media): cuanto más grande sea este valor, más alejados estarían en promedio los puntajes brutos del valor medio; cuanto más pequeño, más cercanos.

El promedio de los puntajes diferenciales se calcula sumando todos y dividiéndolos por la cantidad total de puntajes sumados. Se expresa en la siguiente fórmula.

$$\overline{Pd} = \frac{\sum(PB - \overline{X})}{N} = \frac{\sum Pd}{N}$$

*donde $\overline{Pd}$ es el promedio de los puntajes diferenciales, PB: puntaje bruto $\overline{X}$ media, N: cantidad de puntajes de la muestra.*

El problema que se presenta al calcular el promedio de esta manera es que, por una de las propiedades de la media, la sumatoria de los puntajes diferenciales vale cero, y por lo tanto también su promedio. Observemos éstos tomando los valores de la muestra I.

Tabla 3.14. Muestra I, puntajes brutos y diferenciales, suma y media

| Muestra I | $PB_1$ | $PB_2$ | $PB_3$ | $PB_4$ | $PB_5$ | $PB_6$ | $PB_7$ | $PB_8$ | Σ | X |
|---|---|---|---|---|---|---|---|---|---|---|
| PB | 2 | 3 | 2 | 4 | 17 | 18 | 15 | 19 | **80** | **10** |

| | $Pd_1$ | $Pd_2$ | $Pd_3$ | $Pd_4$ | $Pd_5$ | $Pd_6$ | $Pd_7$ | $Pd_8$ | Σ | X |
|---|---|---|---|---|---|---|---|---|---|---|
| Pd (PB-X) | -8 | -7 | -8 | -6 | 7 | 8 | 5 | 9 | **0** | **0** |

O sea, si se aplica esa fórmula el valor siempre será igual a cero, ya que todos los valores con signo negativo (-) siempre sumarán lo mismo que los de signo positivo (+). En este ejemplo los valores negativos: -8; -7; -8; y -6, suman -29; los positivos: 7; 8; 5 y 9, suman 29; la suma algebraica da igual a cero.

Una de las maneras sencillas de salvar esta dificultad para obtener una medida promedio de los puntajes diferenciales, es elevar al cuadrado los puntajes diferenciales, promediarlos y luego calcular su raíz cuadrada. Al elevar al cuadrado cualquier número se obtiene siempre un número positivo, y hacer un promedio de números positivos asegura que el resultado no podrá ser igual a cero. Es decir, el objetivo de elevar los puntajes diferenciales al cuadrado es justamente aprovechar la propiedad antedicha, hacer desaparecer el signo negativo y con ello evitar que el promedio de los puntajes diferenciales sea nulo. La raíz cuadrada es la operación matemática inversa del cuadrado, es decir que finalmente se aplica para "deshacer" el uso del cuadrado. Matemáticamente esta operación se expresa de la siguiente forma.

$$s = \sqrt{\frac{\sum(pb - \overline{X})^2}{N}} = \sqrt{\frac{\sum Pd^2}{N}}$$



Este estadístico recibe el nombre de ***desvío estándar o desvío típico*** (se lo simboliza con **s, sd, ds ó DE** si se calculó con datos provenientes de una muestra y con σ si se lo calculó con datos de una población), y es la raíz cuadrada del promedio de los puntajes diferenciales elevados al cuadrado. Es una medida que varía de acuerdo a cuan dispersos estén los puntajes brutos respecto de la media. Si el valor es muy pequeño indicará que la muestra está agrupada alrededor de la media; valores más elevados implican mayor lejanía de los PB con respecto al promedio.

En las siguientes tablas se calcula el desvío estándar para las muestras I y II del ejemplo que se sigue.

Tabla 3.15. Muestra I: puntajes diferenciales y su elevación al cuadrado

| Muestra I de 8 casos | | | | | | | | | Σ | $\overline{X}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| PB | 2 | 3 | 2 | 4 | 17 | 18 | 15 | 19 | **80** | **10** |
| Pd | -8 | -7 | -8 | -6 | 7 | 8 | 5 | 9 | **0** | **0** |
| $Pd^2$ | 64 | 49 | 64 | 36 | 49 | 64 | 25 | 81 | **432** | **54** |

$s_I = \sqrt{54} = 7{,}34$

Tabla 3.16. Muestra II: puntajes diferenciales y su elevación al cuadrado

| Muestra II de 8 casos | | | | | | | | | Σ | $\overline{X}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| PB | 8 | 9 | 7 | 8 | 12 | 11 | 13 | 12 | **80** | **10** |
| Pd | -2 | -1 | -3 | -2 | 2 | 1 | 3 | 2 | **0** | **0** |
| $Pd^2$ | 4 | 1 | 9 | 4 | 4 | 1 | 9 | 4 | **45** | **5,62** |

$s_{II} = \sqrt{5{,}62} = 2{,}37$

El valor del desvío estándar de la muestra I es igual a 7,34, valor claramente más elevado que el de la muestra II, que es igual a 2,37. O sea, en promedio, los valores de la muestra I están mucho más alejados de la media que los de la muestra II. En otros términos, los datos de la muestra I están más *dispersos* que los de la muestra II, que es más homogénea.

Si bien el desvío estándar tiene también algunas limitaciones para describir la muestra (como por ejemplo que al igual que la media un solo valor muy alejado lo modifica sensiblemente), es un valor fácil de calcular y sensible a la distribución de los resultados, que ha mostrado en la práctica ser muy útil para usarse como valor estable de referencia.

En síntesis, si en lugar de dividir los puntajes diferenciales por la media se los divide por el desvío estándar, se obtiene un puntaje que aporta más información que los hasta acá han sido referidos. Para el ejemplo del PB=18 se obtendrían puntajes z=1,09 y z=3,3 de acuerdo a si se calculó con la muestra I o II respectivamente. A este puntaje se lo denomina puntaje z y se lo calcula de la siguiente manera.

$z = \frac{Pd}{s}$ = $z = \frac{PB-\bar{X}}{s}$ también se expresa: $z = \frac{X-\bar{X}}{s}$

*Donde z= puntaje z, X es el valor obtenido (el puntaje bruto Pb), $\bar{X}$ es el puntaje medio que se obtiene de una muestra de sujetos comparable con el examinado y s el desvío estándar de esa misma muestra.*

El uso del puntaje z tiene gran difusión en los instrumentos de evaluación psicológica y, por lo tanto, conviene conocer algunas de sus propiedades. Entre las más destacables se encuentran las siguientes.

a) Cuando el PB obtenido en una medición es igual al valor de la media, z valdrá cero.
b) Si z es un valor positivo entonces el puntaje bruto con el que se calculó es mayor a la media, y si es negativo, dicho bruto es menor a la media.
c) Cuando el puntaje z=1, entonces la diferencia *PB-X=s*, es decir, un puntaje z=1 se corresponde con un puntaje bruto ubicado el valor de un desvío estándar por encima de la media; análogamente un puntaje bruto que está a un z=-1 indica que este puntaje bruto se corresponde a un valor ubicado exactamente un desvío estándar por debajo de la media. En extensión de esta propiedad, un puntaje z=2 indica que el puntaje está a dos desvíos estándar encima de la media, un z=3 a tres desvíos, etc. Es decir, que el número z indica cuán alejado o cercano a la media está un puntaje bruto en unidades de desvío estándar.

En síntesis, el puntaje z es, entonces, un puntaje transformado– llamado puntaje estándar– que puede obtenerse a partir de un puntaje bruto, cuando se conocen la media y el desvío estándar de los puntajes obtenidos previamente en una muestra (o población). Este puntaje es un número que puede ser positivo o negativo, e indica cuán cercano o lejano al valor promedio de la muestra está el puntaje bruto.

Si bien el puntaje z ofrece más información para el evaluador que el bruto y diferencial, no está aún claro qué tan alto, bajo o medio es un puntaje z determinado. Por ejemplo, qué tan elevado es un puntaje de z=2, o si pueden hallarse puntajes z de 10,100, ó 1000.

Para tener una dimensión de los valores de los puntajes z, conviene conocer otra propiedad de los puntajes z.

d) "Fuera del intervalo[1] $z = (\bar{X}-ks)$ y $z = (\bar{X}+ks)$están como máximo el $100/k^2$ por ciento de los puntajes de la muestra" (Amón, 1980).

Esta propiedad indica que, fuera de un determinado valor k,-uno por debajo de la media y el mismo número por encima de la media– quedan como máximo una cantidad (porcentaje) de puntajes. Para aclarar esta importante propiedad, ejemplificaremos con intervalos de entre -2 y 2, -3 y 3, -4 y 4 y -5 y 5. El intervalo entre z= -2 y z= 2 indica todos los puntajes z que están ente esos dos extremos, como se muestra en la siguiente tabla.

1. Intervalo es el conjunto de puntajes que hay entre dos de ellos.



Tabla 3.17. Porcentaje de puntajes dentro y fuera de intervalos de z

| Puntajes de z. Entre.... | Porcentaje de puntajes que como **máximo** quedan **fuera** del intervalo | Porcentaje de puntajes que como **mínimo** quedan **dentro** del intervalo |
|---|---|---|
| z=-2 z=2 | 25% | 75% |
| z=-3 z=3 | 11% | 89% |
| z=-4 z=4 | 6,25% | 93,75% |
| z=5 z=-5 | 4% | 96% |

La lectura de la primera fila es: independientemente de cual fuere el instrumento de medición y la muestra en la que este se aplique, como máximo el 25 % de lo individuos que conforman la muestra obtienen puntajes de z menores o iguales a -2 ó z mayores o iguales a 2 , es decir que como mínimo entre los puntajes de z mayores a -2 y menores a 2 se encuentran el 75 % de los datos (puntajes).

De la misma manera, se puede interpretar el cuarto renglón: entre los puntaje z de -5 y 5 están, en el peor de los casos, (cuando las respuestas de los sujetos de la muestra que dan valores de PB muy atípicos e infrecuentes) el 96% de los puntajes obtenidos.

Si los puntajes son más "normales" (ahora en el sentido de comunes, sin puntajes muy atípicos, extremos), estos porcentajes de puntajes brutos incluidos en cualquier intervalo –z + z es muchísimo mayor todavía, por lo cual hay pocos instrumentos donde tenga sentido usar valores de z superiores a 5 o menores a -5, ya que son valores extremos por lo alto o lo bajo.

En el renglón tres puede observarse que entre los valores z= -3 y z= 3, se encuentra, como mínimo, el 89 % de los puntajes. Es decir, que un puntaje z= 3 es un valor sin duda muy elevado y un z = -3 es muy bajo, sea cual fuere el instrumento o la muestra. Cuando la distribución de frecuencias es "normal" (para mayor claridad ver aparado 3.4), ese valor mínimo se eleva a más del 99%), es decir, los puntajes z= 3 ó z= -3 son extremos, muy elevados o muy bajos y mucho más si la distribución de frecuencias es "normal".

De ordinario, en los instrumentos de psicología, la enorme mayoría de los puntajes brutos están a no más de tres desvíos por encima y por debajo de la media, vale decir que valores de z= 3 o menores a z=-3, son casi siempre extremos muy altos o muy bajos, por lo que la gran mayoría de los evaluadores, por convención, para las interpretaciones, consideran z= -1,5 o menor como valor indudablemente bajo y z= 1,5 o mayor como valor indudablemente alto.

En el apartado 3.4. el lector puede consultar cómo son los valores de z y sus frecuencias en la distribución normal.

## Puntaje T

Si bien el puntaje z, comparado con el puntaje bruto, aporta indudables ventajas para la interpretación, el hecho de que sus valores suelan tener decimales y, además, que arrojen puntajes con signo positivo y negativo, complican su facilidad de lectura.

Para simplificar la interpretación de los puntajes estándar, numerosos instrumentos utilizan variantes del z, que pueden calcularse con simples operaciones matemáticas. Todas estas variantes se obtienen sumando una constante para trasladar el valor medio desde el cero hasta un nuevo valor, y multiplicar la puntuación z por otra constante que lo eleva. Con ello se evitan tanto los números negativos, como los decimales.

Una de estas alternativas es el puntaje T –también llamado ***T lineal***– , que se obtiene de la siguiente manera.

$$T = 50 + 10z \qquad T = 50 + \left(\frac{PB - \overline{X}}{s}\right) x10 \quad \text{es decir} \quad T = 50 + \left(\frac{X - \overline{X}}{s}\right) x10$$

*Donde T= puntaje T, z= puntaje z, PB es el puntaje bruto (también X), $\overline{X}$ es el puntaje medio que se obtiene de una muestra de sujetos comparable con el examinado y s el desvío estándar de esa misma muestra*

Nótese que T no es más que el puntaje z al que se lo ha multiplicado por 10 y se le ha sumado un valor de 50. De esta forma, a un valor de z= 0 –que, como ya se ha dicho, se obtiene si el puntaje bruto es igual a la media– se convierte en un puntaje de T= 50.

Un puntaje z=1,- que, como se ha explicado, implica un puntaje bruto de un desvío estándar por encima de la media-, ahora se ha transformado en un puntaje T=60; un puntaje z= -1 equivale al puntaje T=40.

A un puntaje z= -0,6 le corresponderá un puntaje T= 44: nótese que en el nuevo puntaje se ha eliminado el signo menos y la cifra decimal.

La tabla que sigue brinda información sobre las equivalencias que pueden establecerse entre algunos puntajes z y T.

Tabla 3.18. Equivalencias T y z

| Z | -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |

Ya se ha mencionado que en los puntajes z los valores a menos de tres desvíos (z menores a -3 que equivale a T menores a 20) y a más de tres desvíos (z mayores a 3 que equivalen a T mayores a 80) son muy poco frecuentes, por lo que muchos autores, por practicidad, desisten incorporarlos a las conversiones de puntajes de sus instrumentos, asignandoles el puntaje extremo. Por ejemplo, si el constructor de la técnica considera que estos puntajes brutos –cuya transformación a T superan el T=80-, tienen una interpretación psicológica similar a la que hubiera correspondido a un T= 80, podrá optar por asignar a esos puntajes brutos el valor transformado T 80. Análogamente, si los puntajes brutos cuyo puntaje T es menor a 20 puntos, pero su interpretación psicológica es idéntica a la de los que obtuvieron T 20, podrá asignar este puntaje T a esos puntajes brutos. De esta manera se logran perfiles más acotados y fáciles de leer, eliminando zonas de puntajes muy poco frecuentes que no agregan significación psicológica. El test NEO-PI-R (Costa, McCrae 1992; Casullo &, Pérez 2003), el SCL 90-R (Derogatis 1994; Casullo 2007) y el MMPI (Casullo, 1999) son algunos ejemplos de instrumentos que usan este recurso para recortar los puntajes T extremos.



Como al puntaje z= -5 le corresponde un puntaje T= 0, si el puntaje z es menor a -5, el puntaje T que le corresponde será negativo. Sin embargo, como que es muy poco probable obtener puntajes z menores a -5, no es usual encontrar instrumentos que tengan valores de T menores a cero, aunque estos existan en rigor. Análogamente, tampoco es frecuente obtener puntajes z= 5 o mayores, que equivalen a T=100 o más, aunque también, en rigor, esos puntajes existan. Por ejemplo, uno de los pocos instrumentos que expresa sus resultados en puntajes T que llegan hasta los 120 puntos, es el inventario MMPI, en sus dos formas, II y A.

## Puntajes CI

Las escalas Wechsler de inteligencia, para niños y adultos, cuyas versiones actuales son el WISC-IV y el WAIS-III respectivamente (Wechsler 1995, 2002), expresan sus resultados principales en puntajes transformados –también derivados de z-, llamados Coeficientes Intelectuales y Puntajes Índice. En la siguiente tabla se muestran los coeficientes e índices tradicionales.

Tabla 3.19. Coeficientes Intelectuales y Puntajes Índice de las Escalas Wechsler de Inteligencia

| CI | Índices |
|---|---|
| Coeficiente Intelectual Verbal (CIV) | Índice de Comprensión Verbal (ICV) |
| Coeficiente Intelectual de Ejecución (CIE) | Índice de Organización Perceptual (IOP) |
| Coeficiente Intelectual de la Escala Completa (CIEC) | Índice de Velocidad y Precisión (IVP) |
| | Solo WISC III: Índice de Ausencia de Distractibilidad (IAD) |
| | Solo WAIS III Índice Memoria Operativa (IMO) |

Todos ellos comparten las mismas características como puntajes transformados, usando una media de 100 puntos y un desvío estándar de 15.

Con estos valores definidos, la forma de obtener los CI y los Puntajes Índice es análoga al puntaje T, cambiando las constantes de 50 por 100 y de 10 a 15. La fórmula de conversión quedará entonces como sigue.

$$CI = 100 + 15z \qquad I = 100 + 15z$$

*Donde CI= Coeficiente Intelectual (cualquiera de los tres), I=Índice (cualquiera de los cinco) y z = puntaje z*

La lectura tanto de los puntajes CI como de los Índices es análoga a los T y z. En este caso un CI de 100 puntos indica que el examinado ha obtenido un valor medio; un

CI=115 se corresponde a un puntaje z=1, es decir, un desvío estándar por encima de la media, y uno de 130 a un z=2, es decir, dos desvíos encima de la media; lo mismo vale para cualquier Puntaje Índice.

Al igual que en los puntajes T, los valores muy alejados de la media son harto infrecuentes, más si se tiene en cuenta que el Cociente Intelectual tiene una distribución normal (ampliar con la lectura del apartado 3.4). Por este motivo, valores de más de cuatro desvíos por encima y debajo de la media (correspondientes a 40 y 160 puntos respectivamente) son tan poco comunes, que, aunque pueden calcularse, se los suele igualar a los valores extremos.

$\overline{X}$= 100
S= 15
Valor mínimo= 40
Valor máximo= 160
Rango= 120

Con el objetivo de facilitar la lectura de los CI, las Escalas Wechsler proponen rangos o intervalos de interpretación, que se seleccionaron teniendo en cuenta la frecuencia con la que los sujetos pueden obtener distintos valores de CI.

Tabla 3.20. Intervalos de interpretación para las Escalas Wechsler

| CI | 69 y menos | 70-79 | 80-89 | 90-109 | 110-119 | 120-129 | 130 y más |
|---|---|---|---|---|---|---|---|
| Interpretación | Deficiente | Limítrofe | Media baja | Promedio | Media alta | Superior | Muy superior |

Además de estas escalas, existe otro Test de Inteligencia muy difundido, el Test Stanford Binet –hoy en su cuarta versión–, que utiliza también puntajes CI. Estos son similares a los propuestos en los tests de Wechsler, ya que utiliza un valor medio de 100 puntos, pero con un desvío estándar de 16 puntos. Vale decir que la expresión para obtenerlos es la que sigue.

$$CI = 100 + 16z$$

*Donde CI= Coeficiente Intelectual en el Stanford Binet y z = puntaje z*

Cabe aclarar, respecto de esta última técnica, que en sus orígenes estaba destinada solo a niños, y utilizaba un ***cociente intelectual***, es decir que su obtención se calculaba dividiendo la edad mental del sujeto (que era el resultado del test) por la edad cronológica del examinado, y al valor así obtenido se lo multiplicaba por 100. Se trataba de un índice que se obtenía por medio de un cociente, de ahí su nombre, pero su utilización cayó en desuso frente a las ventajas de calcularlo como coeficiente de la manera indicada.



Puntajes Equivalentes

Las escalas Wechsler, ya mencionadas, obtienen sus valores de CI y de Puntajes Índice de una combinatoria de puntajes obtenidos a través de los distintos subtests que las componen. El WISC III, por ejemplo, está conformado por 13 subtests –6 verbales y 7 de ejecución– y el WAIS III por 14 -7 de cada tipo-.

Tabla 3.21. Subtests de las Escalas Wechsler

| WISC III | | WAIS III | |
|---|---|---|---|
| **Subtests Verbales** | **Subtests Ejecución** | **Subtests Verbales** | **Subtests Ejecución** |
| Vocabulario | Construcción Cubos | Vocabulario | Diseño Cubos |
| Analogías | Completamiento de Figuras | Analogías | Completamiento de Figuras |
| Información | Claves | Información | Dígito Símbolo |
| Retención Dígitos | Búsqueda de Símbolos | Dígitos | Búsqueda de Símbolos |
| Aritmética | Composición de Objetos | Aritmética | Rompecabezas |
| Comprensión | Ordenamiento de Historias | Compresión | Ordenamiento de Láminas |
| | Laberintos | Números y letras | Razonamiento con Matrices |

Cada uno de estos subtests puede administrarse en forma independiente de los otros, según la secuencia de administración que indica el manual. Como resultado de la puntuación de cada uno de ellos se obtienen los correspondientes puntajes brutos que se han de convertir en un puntaje transformado que permita su comparación, además de permitir su agrupamiento tanto en los Puntajes Índice o CI antes descritos.

Cada uno de estos puntajes transformados es también una variante del puntaje z llamado ***Puntaje* equivalente** (en algunas versiones en castellano se le llama puntaje escalar; Wechsler, 1999), que se obtiene con una modificación de la fórmula z ya presentada, pero en este caso resulta conveniente que la media sea de 10 puntos y el desvío de 3. Para lograr esto se igualan los puntajes equivalentes a los puntajes z.

$$z = \frac{X-\overline{X}}{S} = \frac{Pe-10}{3}$$

y despejando, resulta

$$Pe = 10 + 3z$$

*Donde z= puntaje z, X es el puntaje bruto, $\overline{X}$ es el puntaje medio que se obtiene de una muestra de sujetos comparable con el examinado, s el desvío estándar de esa misma muestra y Pe el Puntaje Equivalente.*

Como se puede observar, se trata del puntaje z al que se lo multiplica por 3 y se le suma una constante de 10, de modo similar y con los mismos fines que las conversiones ya realizadas con los puntajes T y CI. Cabe aclarar aquí que el usuario de las Escalas Wechsler no necesita realizar este cálculo, ni tampoco el indicado para los CI, ya que el manual del instrumento incluye las tablas de conversión, tablas que han sido obtenidas mediante estas fórmulas y las presentan al lector ya calculadas.

De forma análoga a lo ya expuesto respecto de los puntajes T, en este instrumento se emplea un criterio de "recorte" en los valores extremos superiores e inferiores de los puntajes equivalentes. Así, los valores que están por debajo de 3 desvíos estándar de la media que se corresponde con el puntaje equivalente 1 (si a la media de diez se le restan 3 veces el desvío estándar que vale 3 el resultado es uno: 10- (3X3)=1) se los iguala al valor 1, y todos los puntajes que están por encima de 3 desvíos sobre la media (10+9= 19) se los iguala al valor 19. Debe quedar claro, entonces, que el límite inferior de 1 punto y el superior de 19 se impusieron con un criterio de practicidad, es decir que con la fórmula se pueden obtener mayores y menores a esos límites.

De esta forma, las características de los puntajes equivalentes de las escalas Wechsler pueden resumirse en el siguiente cuadro.

| |
|---|
| $\bar{X}$= 10 |
| S= 3 |
| Valor mínimo= 1 |
| Valor máximo= 19 |
| Rango= 20 |

En los análisis tradicionales de los puntajes equivalentes de los tests de inteligencia o habilidades, un valor por encima de la media en un desvío (o más) se considera elevado, "punto fuerte" o "fortaleza", mientras que a un valor por debajo se lo denomina "punto débil" o "debilidad" (Cayssials,1998).

Por otro lado, en las versiones más recientes del instrumento (WAIS III, WISC IV), el uso de puntos "fuertes" y "débiles" se mantiene pero ya no se lo hace a partir de un valor constante de ±3 puntos para todos los subtests, sino que se debe calcular, escala por escala, el valor que se considera "fuerte" o "débil" para cada subtest o habilidad en cuestión (Kaufman & Lichtenberger, 1999; Wechsler, 1999; Wechsler, 2002 y 2005).

## 3.4 Puntajes y distribución normal

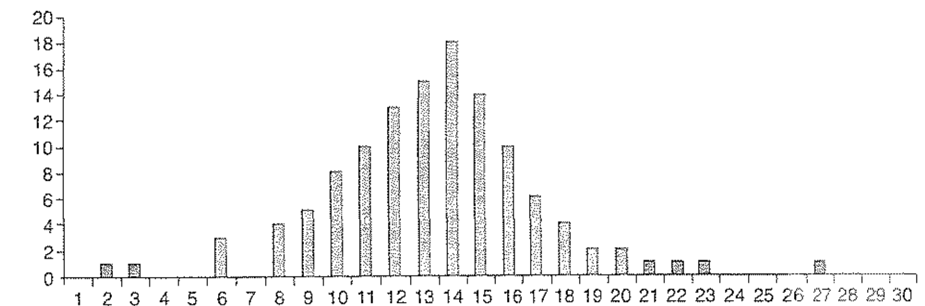
### Distribución normal

Para introducirnos en el análisis de este tipo de distribución de puntajes es necesario repasar la noción de "curva normal". Lo haremos retomando el ejemplo de la escala de memoria (apartado 3.2), aplicada a una muestra de 120 sujetos obteniendo los resultados de la tabla 3.3.

Si se grafican estos datos, las frecuencias, mediante un diagrama de barras, donde la altura de cada una (y) indique el valor de la frecuencia, y el eje horizontal (x) el valor del puntaje, se obtendrá la siguiente gráfica de distribución de frecuencias (también denominada curva de distribución de frecuencias).



Gráfico 3.5.




Nótese que hay una cantidad de frecuencias más elevadas en la parte central y que estas van disminuyendo tanto hacia los extremos de los puntajes inferiores como superiores, es decir, la variable memoria, cuantificada a través de esta técnica, concentra los resultados alrededor del puntaje 14, siendo menor la cantidad de sujetos que puntúan lejos de ese puntaje, ya sea por encima o por debajo del mismo.

Este tipo de distribución de frecuencias es uno de los más comunes, y se presenta en muchísimas medidas, tanto psicológicas como de otro tipo. Los conjuntos de valores productos de mediciones como caracteres morfológicos y fisiológicos de los individuos o de las especies, de magnitudes físicas, de constructos sociológicos y psicológicos, tienen con mucha frecuencia una distribución de esta forma. Tal repetición fundamenta su nominación: *distribución de frecuencias normal*, o en forma abreviada, *distribución normal*.

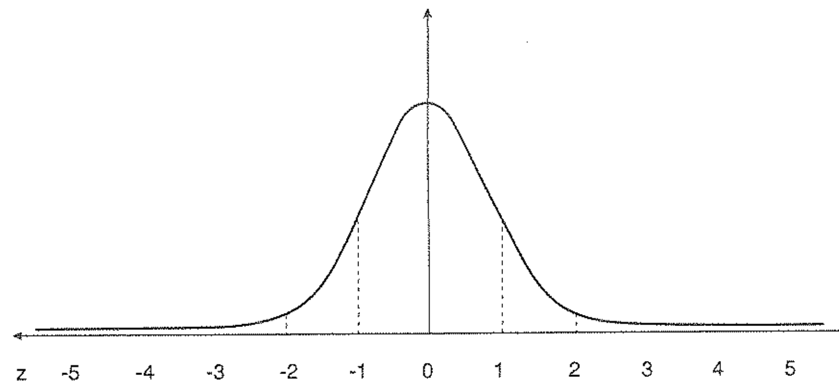
Quien primero describió esta distribución fue Abraham De Moivre, y también trabajó en ella Pierre Laplace, pero fue el matemático, astrónomo y físico alemán Johann Gauss quién formuló un modelo y ecuación para hacer cálculos con distribuciones de este tipo. La ciencia reconoció sus aportes llamando a esta distribución normal como distribución Gaussiana, "curva de Gauss" ó "campana de Gauss".

El modelo de distribución normal propuesto por Gauss parte de varias postulaciones. La variable debe ser continua y debe poder tomar valores infinitamente pequeños y grandes. Si bien algunos de estos supuestos son impracticables en las medidas psicológicas (en rigor en casi todas las medidas), la utilización del modelo de Gauss es más que satisfactorio en muchísimas mediciones psicológicas, y ha demostrado ser en la práctica de enorme utilidad.

En vista de su gran difusión y utilización, a continuación se revisan algunas de sus principales características, dejando al lector la posibilidad de consultar bibliografía específica si necesitara profundizarlos (Amón, 1980; Botella & San Martín,1997; Coolican,1997; Cortada de Kohan, 1994; Pagano,2006).

El siguiente gráfico presenta la curva de distribución normal y más abajo se definen sus características matemáticas.

Gráfico 3.6. Curva de distribución normal




## Características de la distribución teórica normal (modelo matemático)

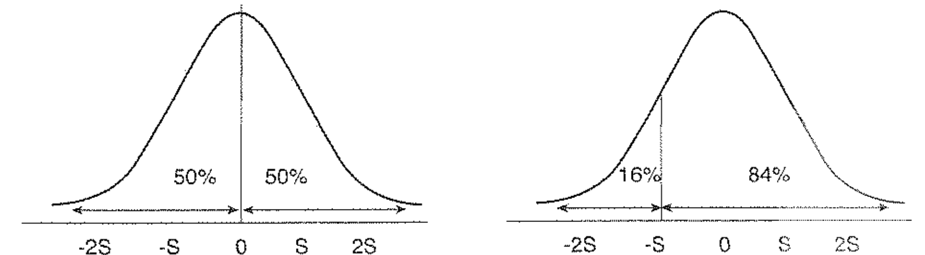
- La altura de la curva en cada punto que la compone representa la densidad de frecuencia, es decir la frecuencia de la variable que se está graficando en la curva dividida la frecuencia máxima. La altura máxima de la curva, representa a la mayor frecuencia dividida por si misma por lo que su valor es uno, y se halla en el centro de la distribución.
- La curva es asintótica tanto hacia los valores mayores como los menores, es decir la altura nunca llega a cero, por ende nunca toca el eje de absisas.
- La media y la mediana coinciden en el mismo valor que se encuentran en el centro de la distribución. La media, al coincidir con la mediana, deja por debajo la mitad de la distribución (50 %) y por encima a la otra mitad. La curva es simétrica.
- Hay dos puntos destacados en la curva (uno por debajo del centro y otro encima) en el cual esta pasa de convexa a cóncava. Estos puntos se corresponden con un desvío estándar por encima o uno por debajo del valor central, del que ya se dijera corresponde a media y mediana.
- Al valor central se le asignó arbitrariamente el valor de cero y al punto que le corresponde al desvío estándar la unidad, es decir el valor 1. Adoptada esta convención, el cero estará en el centro y la unidad de la curva es el desvío estándar.
- Puede demostrarse que para un determinado valor de la variable, el área de la curva que queda hacia su izquierda (en el sentido de los valores menores) representa la frecuencia acumulada que hay por debajo del mismo, es decir, ese área es proporcional a la cantidad de medidas que han obtenido dicho puntaje o puntajes menores. Asimismo, el área de la curva que queda hacia la derecha de ese valor representa la frecuencia de las medidas que han obtenido más de ese valor. Si se divide el área de la izquierda de un determinado valor por el total del área de la curva, y se la multiplica por 100, se obtiene la frecuencia relativa de sujetos que obtuvo esa puntuación o menor, es decir la frecuencia acumulada. Esa propiedad es una de las más usadas de la curva normal, como se verá en el próximo apartado, y por ello es que se ha tabulado para cada puntuación del eje de absisas cuál es la frecuencia acumulada que le correspondería.



- De la misma manera, el área que queda encerrada entre dos valores distintos es proporcional al porcentual de casos o valores que hay entre esos dos puntajes

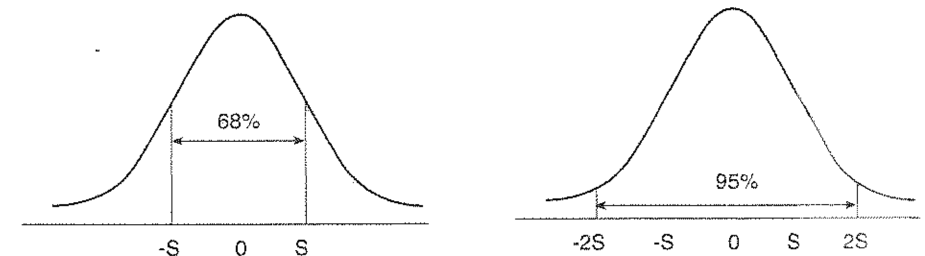
A continuación y a fin de ejemplificar lo antedicho, se muestran graficadas algunas de las áreas de la curva normal más utilizadas en medición psicológica. Los valores porcentuales, salvo el 50 %, fueron redondeados para mayor claridad.

Gráfico 3.7. Porcentaje respecto del total del área para distintos valores de la curva de distribución normal




En el gráfico puede observarse que el área que se encuentra hacia los valores debajo de la media (o mediana) representan el 50 % del total del área, es decir que por debajo de la media quedan el 50 % de los puntajes, lo mismo que por encima de ella.

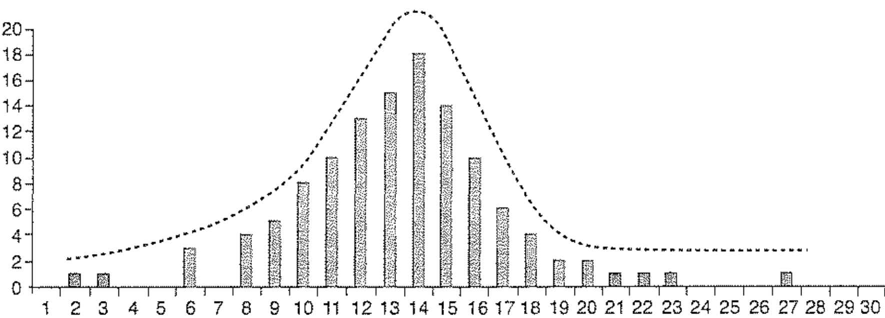
En este otro, puede observarse que el área que se encuentra debajo de la curva en los puntajes que están debajo de un desvío estándar representan aproximadamente 16 % del total del área, es decir que por debajo de un desvío estándar está el 16 % de los puntajes, y por encima de él el resto (84% aproximadamente).




Aquí puede observarse que el área que se encuentra bajo la curva entre un desvío estándar por debajo de la media y uno por encima, representa el 68 % del total del área, es decir que en ese intervalo estaría el 68 % de los puntajes.

El área que se encuentra bajo la curva en el intervalo comprendido entre dos desvíos estándar por debajo y por encima de la media, representa el 95 % del total del área, es decir ese porcentaje de puntajes.

Si una variable psicológica se distribuye de modo suficientemente parecido a la curva normal, es posible entonces trabajarla como si fuera "normal". Por ejemplo, en la gráfica de distribución de frecuencias de la escala de memoria que venimos describiendo, se puede observar cierto parecido a la curva normal: hay muchos puntajes concentrados en valores centrales y cada vez menos en los extremos. La figura siguiente ilustra el "parecido" entre la distribución del ejemplo y la curva normal




Es fácil imaginar que si en lugar de haber evaluado 120 sujetos se hubiesen evaluado miles, el dibujo se asimilaría mucho más a la curva normal teórica. Si se conviene o demuestra con suficiente rigor que la distribución de la variable psicológica en cuestión se aproxima a la de una variable normal, es lícito aplicar las propiedades e instrumentos de este modelo ideal.

## Equivalencias entre medidas estándar y de posición

Cuando la distribución de frecuencias es normal, cada puntaje normalizado (z, T, CI, etc.) dejará por debajo de sí un exacto valor del área de la curva, área que, como se dijo, indica el porcentaje de casos que queda por debajo de ese valor. Ya vimos en los ejemplos que para un valor de un desvío por debajo de la media, que es lo mismo que decir un puntaje z=-1, el porcentaje de casos que quedan por debajo es del 15,87 % y por encima 84,13 %. Como el porcentaje de casos que queda por debajo de un valor se corresponde con el percentil, siempre que la distribución sea normal es posible hallar la correspondencia exacta entre estos y los puntajes estándar.

Se presenta a continuación una tabla donde se pueden ver las equivalencias entre los puntajes z, T, CI de las escalas Wechsler, CI de las escalas Stanford Binet y los percentiles.



Tabla 3.22. Equivalencias z, T, CI y percentil

| Z | T | CI W | CI SB | percentil | Area | Z | T | CI W | CI SB | Percentil | Area |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -4 | 10 | 40 | 36 | 1 | 0,0001 | 0 | 50 | 100 | 100 | 50 | 0,5 |
| -3,9 | 11 | 42 | 38 | 1 | 0,0001 | 0,1 | 51 | 102 | 102 | 54 | 0,5398 |
| -3,8 | 12 | 43 | 39 | 1 | 0,0001 | 0,2 | 52 | 103 | 103 | 58 | 0,5793 |
| -3,7 | 13 | 45 | 41 | 1 | 0,0001 | 0,3 | 53 | 105 | 105 | 62 | 0,6179 |
| -3,6 | 14 | 46 | 42 | 1 | 0,0002 | 0,4 | 54 | 106 | 106 | 66 | 0,6554 |
| -3,5 | 15 | 48 | 44 | 1 | 0,0002 | 0,5 | 55 | 108 | 108 | 69 | 0,6915 |
| -3,4 | 16 | 49 | 46 | 1 | 0,0003 | 0,6 | 56 | 109 | 110 | 73 | 0,7257 |
| -3,3 | 17 | 51 | 47 | 1 | 0,0005 | 0,7 | 57 | 111 | 111 | 76 | 0,758 |
| -3,2 | 18 | 52 | 49 | 1 | 0,0007 | 0,8 | 58 | 112 | 113 | 79 | 0,7881 |
| -3,1 | 19 | 54 | 50 | 1 | 0,001 | 0,9 | 59 | 114 | 114 | 82 | 0,8159 |
| -3 | 20 | 55 | 52 | 1 | 0,0013 | 1 | 60 | 115 | 116 | 84 | 0,8413 |
| -2,9 | 21 | 57 | 54 | 1 | 0,0019 | 1,1 | 61 | 117 | 118 | 86 | 0,8643 |
| -2,8 | 22 | 58 | 55 | 1 | 0,0026 | 1,2 | 62 | 118 | 119 | 88 | 0,8849 |
| -2,7 | 23 | 60 | 57 | 1 | 0,0035 | 1,3 | 63 | 120 | 121 | 90 | 0,9032 |
| -2,6 | 24 | 61 | 58 | 1 | 0,0047 | 1,4 | 64 | 121 | 122 | 92 | 0,9192 |
| -2,5 | 25 | 63 | 60 | 1 | 0,0062 | 1,5 | 65 | 123 | 124 | 93 | 0,9332 |
| -2,4 | 26 | 64 | 62 | 1 | 0,0082 | 1,6 | 66 | 124 | 126 | 95 | 0,9452 |
| -2,3 | 27 | 66 | 63 | 1 | 0,0107 | 1,7 | 67 | 126 | 127 | 96 | 0,9554 |
| -2,2 | 28 | 67 | 65 | 1 | 0,0139 | 1,8 | 68 | 127 | 129 | 96 | 0,9641 |
| -2,1 | 29 | 69 | 66 | 2 | 0,0179 | 1,9 | 69 | 129 | 130 | 97 | 0,9713 |
| -2 | 30 | 70 | 68 | 2 | 0,0228 | 2 | 70 | 130 | 132 | 98 | 0,9772 |
| -1,9 | 31 | 72 | 70 | 3 | 0,0287 | 2,1 | 71 | 132 | 134 | 98 | 0,9821 |
| -1,8 | 32 | 73 | 71 | 4 | 0,0359 | 2,2 | 72 | 133 | 135 | 99 | 0,9861 |
| -1,7 | 33 | 75 | 73 | 4 | 0,0446 | 2,3 | 73 | 135 | 137 | 99 | 0,9893 |
| -1,6 | 34 | 76 | 74 | 5 | 0,0548 | 2,4 | 74 | 136 | 138 | 99 | 0,9918 |
| -1,5 | 35 | 78 | 76 | 7 | 0,0668 | 2,5 | 75 | 138 | 140 | 99 | 0,9938 |
| -1,4 | 36 | 79 | 78 | 8 | 0,0808 | 2,6 | 76 | 139 | 142 | 99 | 0,9953 |
| -1,3 | 37 | 81 | 79 | 10 | 0,0968 | 2,7 | 77 | 141 | 143 | 99 | 0,9965 |
| -1,2 | 38 | 82 | 81 | 12 | 0,1151 | 2,8 | 78 | 142 | 145 | 99 | 0,9974 |
| -1,1 | 39 | 84 | 82 | 14 | 0,1357 | 2,9 | 79 | 144 | 146 | 99 | 0,9981 |
| -1 | 40 | 85 | 84 | 16 | 0,1587 | 3 | 80 | 145 | 148 | 99 | 0,9987 |
| -0,9 | 41 | 87 | 86 | 18 | 0,1841 | 3,1 | 81 | 147 | 150 | 99 | 0,999 |
| -0,8 | 42 | 88 | 87 | 21 | 0,2119 | 3,2 | 82 | 148 | 151 | 99 | 0,9993 |
| -0,7 | 43 | 90 | 89 | 24 | 0,242 | 3,3 | 83 | 150 | 153 | 99 | 0,9995 |
| -0,6 | 44 | 91 | 90 | 27 | 0,2743 | 3,4 | 84 | 151 | 154 | 99 | 0,9997 |
| -0,5 | 45 | 93 | 92 | 31 | 0,3085 | 3,5 | 85 | 153 | 156 | 99 | 0,9998 |
| -0,4 | 46 | 94 | 94 | 34 | 0,3446 | 3,6 | 86 | 154 | 158 | 99 | 0,9998 |
| -0,3 | 47 | 96 | 95 | 38 | 0,3821 | 3,7 | 87 | 156 | 159 | 99 | 0,9999 |
| -0,2 | 48 | 97 | 97 | 42 | 0,4207 | 3,8 | 88 | 157 | 161 | 99 | 0,9999 |
| -0,1 | 49 | 99 | 98 | 46 | 0,4602 | 3,9 | 89 | 159 | 162 | 99 | 0,9999 |
| 0 | 50 | 100 | 100 | 50 | 0,5 | 4 | 90 | 160 | 164 | 99 | 0,9999 |

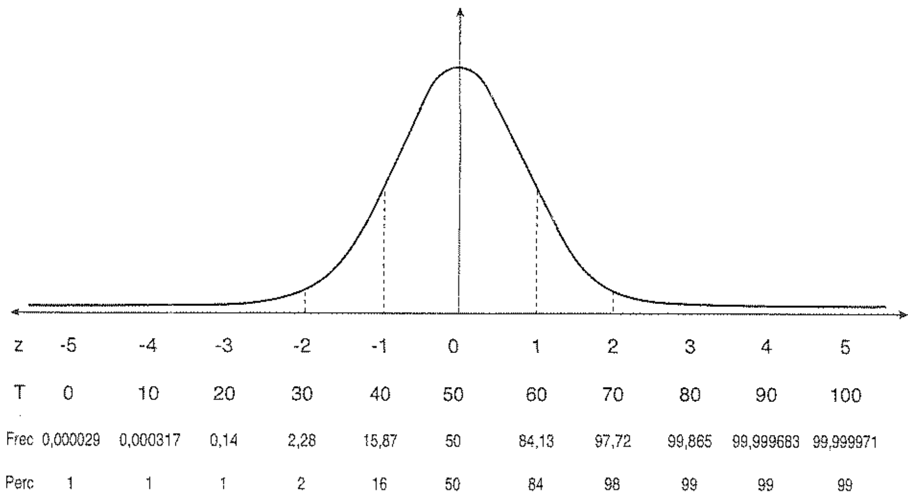
Téngase en cuenta que esta tabla solo muestra la equivalencia numérica entre esos puntajes, lo cual no implica que sea correcta, por ejemplo, la lectura de un puntaje z obtenido en una escala de personalidad en términos de coeficiente intelectual (CI).

Observando los valores de la tabla anterior, se pueden destacar algunas relaciones que refuerzan la comprensión de algunos tópicos vistos anteriormente.

- Los percentiles correspondientes a los puntajes z menores a -3 (ó T=20) se corresponden con el percentil 1, de la misma forma que para los puntajes z mayores a 3 (ó T =80) coinciden con el percentil 99. Estos valores se corresponden al mínimo y máximo de la escala percentilar, es decir son valores muy extremos por lo bajo o por lo alto. Por ello en escalas que evalúan variables psicológicas distribuidas en forma normal, no suele usarse z menores a -3 (o T menores a 20) ni mayores a 3 (o T mayores a 80) ya que es muy poco probable que los sujetos los obtengan (menos del 1% en cualquiera de los dos).
- Del mismo modo, los percentiles correspondientes a los CI con valores de 70 y 130 dejan solo un 1% de puntajes por encima y por abajo respectivamente, marcando prácticamente los extremos de la escala. En el caso de las personas que obtienen puntajes por debajo, se hace patente una gran limitación intelectual, y la lectura inversa es adecuada para las personas que puntúan por encima de 130: esto justifica los intervalos propuestos para el análisis del CI vistos en tabla 3.20.
- En la tabla de equivalencia presentada solo se incluyeron valores desde z=-4 hasta z=4, que como se observa en su equivalente percentilar son valores muy extremos. Cuando una escala de evaluación utiliza en sus resultados valores por encima de z=4 o menores a z=-4, es muy probable que los puntajes de la variable que evalúa ese instrumento no presenten una distribución normal; caso contrario sería poco práctico hacerlo.

En la gráfica siguiente se pueden ver algunas de las equivalencias expresadas en la tabla anterior.

Gráfico 3.8. Equivalencia de puntajes en variables con distribución normal






Comparación de escalas con distribución no normal

Si se calculan los percentiles que corresponden a cada puntaje normalizado en escalas que evalúan variables psicológicas que no se distribuyen según la curva normal, se observará que ya no coinciden con los que les hubiera correspondido si la variable fuera normal; menos coincidencia habrá cuanto más se aparte la distribución de frecuencias de la distribución normal. Esto sucede corrientemente en escalas que evalúan dimensiones psicopatológicas, ya que estas no suelen ser variables con distribución normal.

Algunos instrumentos tratan de medir varias dimensiones de características no normales por medio de un conjunto de escalas, como puede ser el caso de los inventarios de personalidad de Minnesota, MMPI-2 y MMPI-A (Casullo 1999, Casullo 2003). En tales casos, siendo escalas que dan cuenta de constructos clínicos– es decir no normales-, el mismo puntaje estándar puede tener una significación muy distinta si se obtuvo en una escala que en otra.

Por ejemplo, el inventario MMPI–2, al igual que el MMPI-A, constan de diez escalas clínicas y cada una de ellas arroja puntajes T.

Tabla 3.23. Escalas de MMPI-2 y MMPI-A

| | | | |
|---|---|---|---|
| Hs | Hipocondriasis | Pa | Paranoia |
| D | Depresión | Pt | Psicastenia |
| Hy | Histeria | Sc | Esquizofrenia |
| Pd | Desviación Psicopática | Ma | Manía |
| Mf | Masculinidad Feminidad | Si | Introversión Social |

Si bien cada una de las diez escalas arroja un puntaje transformado, el hecho de que cada una de ellas presente distinta distribución de frecuencias, hace que un mismo puntaje no signifique lo mismo en una escala y en otra. Por ejemplo, un puntaje T de 65 en la escala Desviación Psicopática no tiene la misma implicancia que un T de 65 en la escala Histeria. Cabe aclarar que la diferencia de implicancia que se indica no es por la gravedad de los síntomas de uno u otro constructo, sino al hecho de que un valor T de 65 en Histeria implica un percentil distinto que el mismo T en Desviación Psicopática, y esto es producto de la diferencias en la distribución de frecuencias en cada escala.

De modo general, esta dificultad se presenta para valorar los puntajes T de cada una de las diez escalas clínicas del MMPI, ya que cada una tiene una distribución de frecuencia distinta, y por ello si una escala tiene un puntaje T más elevado que otra, en términos percentilares podría estar sucediendo lo inverso: esto dificulta la valoración y la lectura de estos perfiles.

Esta dificultad para interpretar perfiles realizados con variables de distribuciones de frecuencia muy distintas entre sí (y no normales) se suele subsanar de dos maneras. La primera de ellas, es interpretando el puntaje escala por escala e integrando la información a posteriori; esto requiere un conocimiento profundo del significado del puntaje en cada escala, por lo que suele hacerse con el auxilio de un manual o texto guía, e implica buena habilidad del evaluador. La segunda forma, es haciendo alguna modificación de estos puntajes para que sean más comparables.

Esta última opción es la que da lugar a los tres puntajes transformados que se desarrollan a continuación y cuyo objetivo es "normalizar", homogenizar o equiparar resultados en escalas con distribuciones de frecuencias que originalmente tienen una distribución muy distinta de la normal.

### Puntaje T normalizado

Una de las variantes del puntaje T es el denominado puntaje T normalizado, utilizado, por ejemplo, en el SCL 90-R (Derogatis, 1994; Casullo 2007). Se obtiene "normalizando" la distribución de frecuencias de cada una de las escalas de las nueve dimensiones primarias que lo conforman.

Tabla 3.24. Dimensiones de SCL-90-R

| | | | |
|---|---|---|---|
| SOM | Somatizaciones | HOS | Hostilidad |
| OBS | Obsesiones y compulsiones | FOB | Ansiedad fóbica |
| SI | Sensibilidad Interpersonal | PAR | Ideación paranoide |
| DEP | Depresión | PSIC | Psicoticismo |
| ANS | Ansiedad | | |

Una manera práctica de hacerlo es partiendo de la distribución de frecuencias de la curva normal y de su vínculo con los percentiles. Como se resume en tabla 3.22, para la distribución normal hay un valor percentilar exacto que corresponde a cada puntaje T.

Básicamente, para construir los puntajes normalizados, para cada una de las escalas (en el caso del SCL 90 –R para la nueve expresadas en tabla 3.24) se procede del siguiente modo:

1) A partir de cada uno de los puntajes brutos se calcula el percentil o frecuencia acumulada que le corresponde.
2) Se le asigna a ese percentil el puntaje T que le correspondería de acuerdo a la curva normal.

Vale decir entonces, que no se utiliza la fórmula analizada con anterioridad en puntajes T lineales, sino que se utiliza el percentil que se le asigna al puntaje T de acuerdo a la curva normal. De esta forma, por ejemplo un puntaje T normalizado de 60 en cualquier escala corresponde al percentil 84, un puntaje T de 50 al 50, uno de 40 al 16, y esto independientemente de cómo sea la distribución de frecuencias de esa escala.

Esta operatoria tiene la ventaja de que, en términos percentilares, los puntajes T de todas las escalas significan lo mismo, pero esto no debe confundirse con que realmente la distribución sea normal, ni tampoco con que la "gravedad" que implican sea la misma. Así, por ejemplo, en la SCL 90-R un puntaje T= 60 en la escala Depresión implica que el examinado ha superado en esa dimensión al 84 % de los sujetos de la muestra, lo mismo que un T=60 en Hostilidad implicaría que superó al 84 % de la muestra en esa otra dimensión, pero nada nos dice sobre si desde un punto de vista psicopatológico tener 60 en Depresión amerita mayor atención o no que tener 60 en Hostilidad.



En el siguiente cuadro se sintetizan las características del puntaje T estandarizado.

> Valor medio: 50
> Desvío estándar= 10
> A cada valor de T le corresponde el percentil de una distribución normal
> En el caso del SCL 90 se limitó el rango entre puntajes T de 20 a 80

La mayoría de los instrumentos que utilizan estos puntajes permiten obtener un perfil gráfico que expresa los puntajes del sujeto en cada escala. Con el objetivo de que estas gráficas tengan una escala adecuada para poder ser dibujadas, y teniendo en cuenta que los puntajes T muy extremos – ya sean lineales o normalizados– no aportan ya mayor significación, es que se suelen recortar los puntajes T muy elevados o muy bajos (como ya se explicó en el apartado de puntaje T). En el caso del SCL 90 se "recortaron" los extremos del perfil en 20 y 80, y en caso del MMPI-2, los recortes fueron en 30 y 120. En ambos instrumentos los valores cercanos a 1,5 desvíos estándar por encima de la media, se consideran significativos desde el punto de vista interpretativo (puntajes T de alrededor de 65 puntos).

### Puntaje T uniforme

El puntaje T uniforme es otra variante del T lineal, propuesta por Tellegen (Tellegen, 1988; Tellegen & BenPorath, 1992) y es utilizado tanto en el MMPI-2 como en el MMPI-A, para ocho de las diez escalas clínicas : Hipocondriasis (Hs), Depresión (D), Histeria (Hy), Desviación Psicopática (Pd), Paranoia (Pa), Psicastenia (Pt), Esquizofrenia (Sc) y Manía (Ma). Fueron exceptuadas las escalas Masculinidad/Feminidad (Mf) e Introversión Social (Si) por tener distribuciones de frecuencia y forma de construcción distintas de las otras.

Técnicamente se procedió a hacer una distribución de frecuencias "promedio" de estas 8 escalas y con ella calcular los nuevos puntajes T. La forma de construir estos puntajes en este instrumento, es, resumidamente, la siguiente.

a) Se calculan los valores T lineales para todos los puntajes brutos de estas 8 escalas para varones y mujeres, obteniéndose, entonces, 16 conjuntos de puntajes.
b) Para cada valor percentilar, se calculan los 16 puntajes T lineales que corresponden en las 16 distribuciones y se promedian. De esta forma, para cada percentil se obtiene un puntaje T que no es el T lineal de ninguna de las 16 distribuciones, sino el promedio de los 16.
c) Con este conjunto de puntajes T asociada a cada percentil, se hicieron las tablas de conversión de puntaje T-percentil que se usará para las 8 escalas. A ese puntaje T se lo llamó uniforme
d) Finalmente con las tablas de conversión se hace el proceso inverso para cada una de las 16 distribuciones, obteniendo los puntajes brutos asociados a cada T uniforme. Esto se hace con corrección de K- factor corrector que tiene esa técnica– y sin ella y con ellos se confecciona el perfil.

Lo que se logra con este procedimiento, es que la interpretación de las escalas– desde un punto de vista percentilar y de la distribución– es más uniforme. También es más pareja la cantidad de puntajes brutos asignados a cada T.

Se recuerda que el puntaje T uniformado solo se usa en las escalas mencionadas; el resto de las escalas del MMPI-2 y MMPI-A utilizan los tradicionales puntajes T lineales (los que se obtienen con la fórmula ya vista en el apartado correspondiente).

Las principales características del puntaje T uniforme son las siguientes.

Valor medio: 50
Desvío estándar= 10
A cada valor de T le corresponde un percentil similar que el de las otras escalas que lo utilizan
En el MMPI-II el rango se limitó de T 30 a T 120

## Puntajes de prevalencia (pp)

Los puntajes llamados "de prevalencia", propuestos en los instrumentos de Millon -el Inventario Clínico Multiaxial de Millon, hoy en su tercera versión, MCMI-III (Millon 1994), o el Inventario de Estilos de Personalidad de Millon MIPS (Millon, 1997), entre otros del mismo autor-, usan un puntaje que es una variante que combina la distribución de frecuencias con tasas de prevalencias poblacionales de los constructos a evaluar (las escalas referidas evalúan trastornos de la personalidad y estilos de personalidad, respectivamente).

Se usa el término tasa de prevalencia o, simplemente, prevalencia para indicar la frecuencia (generalmente relativa y porcentual) que tiene determinado constructo en una población, es decir qué proporción o porcentaje de sujetos poseen ese rasgo (o ese trastorno) en el total de la población. En el caso de los Trastornos de Personalidad, las tazas suelen estar ya relevadas por estudios epidemiológicos; así, podemos observar que, por ejemplo, en el DSM IV (1995) se indican las prevalencias de varios trastornos.

En el caso de las dimensiones de personalidad, se realizaron estudios actuariales para determinar las prevalencias (Millon, 1997; Castro Solano; Casullo & Pérez, 2004). Una vez hallados esos valores, los llamados puntajes de prevalencia son relativamente sencillos de obtener y tienen como finalidad –al igual que los normalizados– facilitar la lectura e interpretación de los perfiles.

Como los puntajes de prevalencia de las distintas técnicas mencionadas difieren en sus valores de corte y rangos, se ejemplificará con el MIPS el concepto del diseño de estos puntajes. Resumidamente, la forma técnica de realizarlo sigue los pasos siguientes.

a) Se postula el rango que tendrán los valores transformados y un punto de corte a partir del cual se considera que el rasgo está presente (o el trastorno, en el caso de un instrumento clínico). En el MIPS se decidió tomar un rango de 0 a 100 puntos y como punto de corte a partir del cual se considera que está presente el rasgo al valor 50. Es importante destacar que estos valores se eligen por convención y no están relacionados con, por ejemplo, los puntajes percentilares.
b) Al tomar esa decisión, se ha resuelto que todas las personas que no poseen el rasgo deberán puntuar por debajo de 50, y todas las que si lo poseen, por encima de este valor. Es decir, el puntaje de prevalencia, desde esta óptica, es categorial: a partir del valor 50 hacia arriba se dará una medida de la presencia y por debajo de 49 se dará una medida de la ausencia del mismo constructo.



c) Si la prevalencia de una determinada dimensión es, por ejemplo, del 65% (el sesenta y cinco por ciento de la población se reconoce con ese rasgo), entonces el 65 % de los sujetos deberá obtener en el instrumento un puntaje de prevalencia (pp) por encima de 50 (posee el rasgo), y el 35 % puntuará por debajo de ese valor (no posee el rasgo). Para lograrlo se obtienen los valores percentilares correspondientes a los puntajes brutos, y al puntaje bruto que corresponde al percentil 35 se le asignará el puntaje pp de 50, garantizando, de esta forma, que el 65% de los sujetos obtengan 50 o más puntos de pp.
d) Todos los puntajes brutos que caen en percentiles superiores al percentil 35, se distribuirán arriba del pp 50, siguiendo pautas de distribución basadas en rangos percentilares, y en los puntajes brutos que quedaron por debajo del percentil 35, se hará lo propio por debajo del 50. Para hacer esta asignación de puntajes brutos a pp, en el caso particular del MIPS, se eligieron valores de referencia: se asignó el pp 69 a la mediana de la distribución de los que se reconocieron con el rasgo (percentil 50), y al valor 89 al percentil equivalente a un desvío estándar de la curva normal (percentil 84). Se utilizó el mismo criterio de distribución hacia abajo del pp 50, siendo el 29 el valor correspondiente a la mediana de los sujetos que no se reconocieron con el rasgo, y el pp 9 el equivalente al percentil de un desvío estándar hacia abajo (16%). Todos los puntajes intermedios se asignaron siguiendo intervalos uniformes.

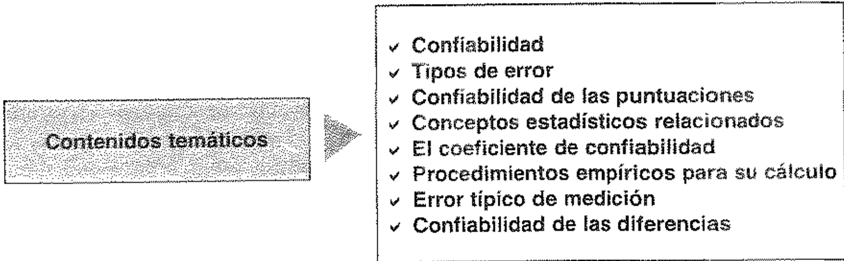
Los puntajes de prevalencia obtenidos de esta forma quedan, entonces, caracterizados de la siguiente manera.

Rango: 0-100 puntos.
Punto de corte (ausencia- presencia de rasgo):49: Esto implica que con cincuenta o más puntos el rasgo de que se trate se encuentra "presente".
*Bandas en presencia de rasgo:*
Entre pp: 50-69 puntuaría el 50% de la gente con el rasgo presente: indica que lo caracteriza.
Entre; 70-89 puntuaría el siguiente 34 % de sujetos que se caracterizan por tener más conductas prototípicas del rasgo, más exacerbado o más frecuente)
Entre pp: 90-100 puntuaría el 16% con el rasgo prototípico
*Bandas de ausencia de rasgo: estas bandas se usan para contrastar los valores de dos escalas bipolares antitéticas. Cuando una de ellas dio su puntaje por encima de 50 (presencia de rasgo) se observa la escala complementaria, y si esta da por debajo de 50 se usan las siguientes bandas para ver el contraste con la primera.*
Entre pp: 49 y 30 puntuaría el 50% de la gente.
Entre pp: 29 y 10 puntuaría el 34% de la gente.
Entre pp: 9 y 0 puntuaría el 16% restante.

*Capítulo 4*

# *Confiabilidad y error de medición*

Alicia Cayssials




## 4.1 Confiabilidad

El presente capítulo desarrolla contenidos específicos que, a su vez, conservan fuertes nexos con los que se exponen en el anterior y con los que se tratan en el siguiente. Conceptualmente, los tres capítulos se refieren a las puntuaciones obtenidas a través de una técnica psicométrica. En el precedente, se analizan los distintos tipos de puntajes considerados básicos en psicometría, mientras que en los que le siguen, se pone el foco en las cualidades de los mismos, en sus bondades y en su ajuste a los requerimientos científicos. Debido a que la elaboración de una técnica psicométrica implica, no sólo seleccionar adecuadamente el tipo de puntuaciones que el instrumento ha de brindar, es necesario también proveer evidencia empírica sobre la calidad de las mismas, lo cual se lleva a cabo poniendo a prueba la consistencia y precisión de los puntajes –*confiabilidad*– (tema a desarrollar en el presente capítulo), y analizando la bondad de dichos puntajes para aportar información pertinente de la variable que se intenta medir –*validez*– (cap. 2).

Específicamente, la *confiabilidad* es un índice de la calidad de la técnica de evaluación, que, si bien todo usuario debe poder comprender y saber valorar, su indagación está a cargo de quien elabora o adapta el instrumento en cuestión. En psicometría se la estudia de un modo técnico y cuantitativo, de ahí que se ha destinado el apartado 4.4. a repasar algunos conceptos estadísticos relacionados con esta temática.

A modo de introducción al tema de la *confiabilidad* de las puntuaciones, es necesario destacar que toda medición científica se halla fundamentada en una Teoría de la Medición, la cual analiza distintas propuestas para describir, categorizar y evaluar la calidad de las medidas, y que, a su vez, tiene como objetivos tanto mejorar su utilidad y su precisión, como desarrollar nuevos métodos en la obtención de instrumentos de

mayor calidad. En este capítulo -así como en el resto de este libro-, la medición en Psicología se analiza desde la óptica de la llamada *Teoría Clásica de los Tests* (TCT), denominada también *Modelo del Valor Esperado*.

La TCT es una teoría útil para describir la influencia de los errores de medida en las puntuaciones observadas u obtenidas a través de instrumentos, y sus relaciones con las puntuaciones verdaderas. Se basa en el Modelo Lineal de Spearman, desarrollado a principios del siglo XX. Se trata del primer modelo que aborda el problema de la incertidumbre o error inherente a cualquiera de las medidas realizadas mediante la aplicación de un test. Aún cuando posteriormente se han desarrollado nuevas teorías -por ejemplo, la Teoría de la Respuesta al Ítem (TRI) y sus variantes-, la TCT continúa vigente en la actualidad.

Si se acepta la posibilidad de medir en Psicología, es necesario -según la TCT-, asumir dos supuestos:

- Existen puntajes verdaderos, que reflejan puntualmente la realidad, que miden de un modo exacto, sin error.
- Siempre que se realizan mediciones pueden cometerse errores.

Estos supuestos pueden parecer contradictorios. Sin embargo, si se tiene en cuenta que el primero es un supuesto ideal -una hipótesis de trabajo, una probabilidad teórica-, esta contradicción desaparece. Se supone, entonces, la existencia de *puntuaciones verdaderas*, sin error y al mismo tiempo se supone que, al realizar una medición concreta del fenómeno o atributo en cuestión, lo más probable es que se cometan errores.

La distinción entre un puntaje teorizado, ideal, que llamaremos *verdadero*, y otro concreto, que llamaremos *obtenido* -el que resulta de la aplicación de una técnica psicométrica-, es fundamental ya que uno de los objetivos más importantes de la Psicometría es determinar el valor real o *puntuación verdadera*. Esta puntuación se define como lo que queda de la puntuación *observada* u obtenida a través de un test, una vez eliminados los errores de media. Podemos formalizar este enunciado en la siguiente fórmula.

$X = X_v + X_e$ donde despejando

$$X_v = X - X_e \qquad [1]$$

Donde $X_v$ es el puntaje verdadero, hipotetizado, ideal; que es igual al **X** -puntaje obtenido a través de una técnica-, al que hay sustraerle el $X_e$, el puntaje debido al error, el error que es probable que se esté cometiendo al medir.

Los mismos conceptos se presentan en la siguiente figura, donde el círculo completo representa al puntaje obtenido a través de la aplicación de una técnica, y se lo muestra dividido, descompuesto, en dos porciones, una que grafica el puntaje verdadero y la otra el puntaje de error.



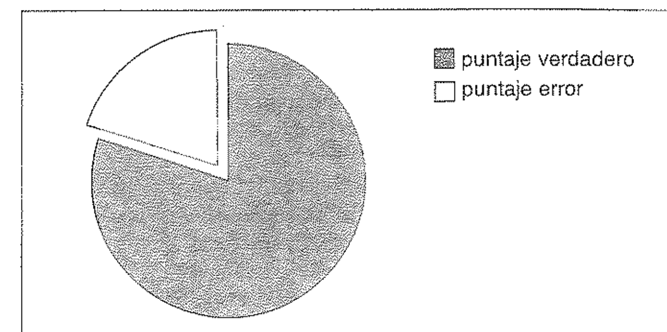



Figura 4.1. Representación gráfica de la composición del puntaje obtenido.

Se han elaborado modelos formales o teorías de los tests que permiten analizar las relaciones entre estos tres componentes básicos: las puntuaciones *observadas*, los *errores* de medida y las puntuaciones *verdaderas*.

En las mediciones indirectas de fenómenos intangibles - como son la mayoría de las que se realizan en psicología- el *puntaje verdadero* no puede ser concretamente calculado, dado su carácter ideal (ver apartado 3.1.). O sea, se trata de un concepto al que se intenta acceder y del cual sólo puede inferirse su valor hipotético. Por esta razón, el objetivo de los estudios que se realizan sobre la precisión de las puntuaciones obtenidas a través de un instrumento es controlar y calcular el margen de error.

En cuanto a la organización del capítulo, comienza con el abordaje de las distintas fuentes de error en las puntuaciones obtenidas a través de una técnica de evaluación psicológica; luego propone un repaso de los conceptos estadísticos elementales y fundamentales en el estudio de la *confiabilidad* y su tema complementario, el *error de medición*. Más adelante, retoma el tema de la confiabilidad, para reconsiderarlo a la luz de los distintos diseños y procedimientos empíricos que permiten calcular el *coeficiente de confiabilidad*. Por último, vuelve al análisis del *error* para referirlo, de modo concreto, al *error de medición* y a la *confiabilidad* en la valoración de las *diferencias entre puntajes* obtenidos, ambas cuestiones de suma importancia y utilidad en la práctica profesional.

## 4.2 Tipos de error

Una medición, realizada a través de un instrumento psicométrico, puede estar influida por fuentes de error tan diferentes y variadas que no es posible mencionarlas exhaustivamente. Sin embargo, se las puede dividir en dos categorías según el tipo de error que generan: *sistemáticos* o *asistemáticos*. A su vez, el estudio de cada uno de estos tipos de error implica llevar a cabo distintos estudios psicométricos para su estimación y control. Dedicaremos un apartado a cada uno de ellos ya que es fundamental que el lector los distinga claramente.

### Errores sistemáticos

Las fuentes de errores sistemáticos son aquellas que desplazan las puntuaciones en cierta dirección, generando una puntuación sistemáticamente elevada o baja. Suelen denominarse también errores constantes (Hogan, 2003).

Concretamente, para ilustrar este punto, vamos a recurrir a un ejemplo que se encuentra con frecuencia en la literatura psicométrica para introducir al tema y que ha resultado útil. Se trata de "la balanza" (que viene a sustituir al test). Si una "balanza", consistentemente, en todas las ocasiones, indica 1,5kg más (o menos, no importa esto) a todos los sujetos que se pesan en ella, obviamente, no pesa bien, es decir, no indica "el peso verdadero". Nótese también que, si se evalúa con esta balanza no sólo a un sujeto sino a una *muestra* de sujetos, el peso relativo de las personas permanece sin cambios, entendiendo como tal la diferencia de pesos entre ellos. Dicho de otro modo, la circunstancia de que una "balanza" pese sistemáticamente 1,5kg más (o 1,5kg menos) a todos los que se pesan en ella incluirá una *constante* que se suma (o se resta) al valor verdadero en cada medición. Se trata de un ejemplo que viene a mostrar que el error sistemático, a pesar de introducir diferencias en el resultado de la medición, no cambia la variabilidad, la distribución de las puntuaciones de un grupo de sujetos en la variable que se está evaluando. Obviamente, una balanza con estas características llevaría a errores en el cálculo del IMC[1] y estos errores serían trasladados al diagnóstico y a las recomendaciones posteriores. Los instrumentos que conllevan este tipo de error sistemático sobreestiman (o subestiman, según el caso) el atributo evaluado (peso, en este ejemplo).

Tomemos ahora un par de ejemplos donde se consideren variables psicológicas: un test de inteligencia y una técnica que evalúa depresión. En el primer caso tendríamos un test que sistemáticamente sobrevalora (o subvalora) el nivel de inteligencia de los sujetos y en el segundo estos errores influirían (en un sentido o en otro), en la evaluación del nivel de depresión de las personas examinadas. Estos errores sistemáticos ocasionarían a su vez errores en el diagnóstico intelectual o psicopatológico, respectivamente.

Los errores sistemáticos pueden ser detectados a través del análisis de la *validez* del instrumento. Los estudios sobre la confiabilidad se ocupan de los errores *asistemáticos*, a los que se dedica el siguiente apartado.

### Errores no sistemáticos

Los errores no sistemáticos, también llamados asistemáticos, son aquellos sin posibilidad de ser controlados, impredecibles o *aleatorios*, ya que son generados por las variaciones cuya causa es el azar.

Las técnicas psicométricas son instrumentos estandarizados, lo cual implica, entre otras cuestiones, uniformidad en el proceso de administración y de evaluación, ya que las variaciones en los procedimientos darían lugar a variaciones en las respuestas no atribuibles a la variable que se desea medir (véase cap. 1.3). Sin embargo, aunque las pautas de aplicación y evaluación, señaladas por el autor en el Manual de la técnica a aplicar, sean respetadas, al llevar a cabo una medición siempre existen factores o condiciones azarosas que pueden generar errores. Dichas fuentes de error pueden haber sido generadas en la etapa de construcción de la técnica, en la administración, en la puntuación y en la interpretación de los resultados arrojados por la misma (Cohen y Swerdlik, 2000).

1. El **Índice de Masa Corporal** (IMC) es uno de los métodos más utilizados para el diagnóstico de obesidad, y a su vez, para establecer el grado de obesidad. Se obtiene dividiendo el peso en kilos por la estatura en metros elevada al cuadrado (kg/m2).



En cuanto a las fuentes que pueden generar errores durante la etapa de construcción de un instrumento, una de las posibles es la del *muestreo de contenido*, que se refiere a la variación en los resultados obtenidos dependiendo de los ítems incluidos en la técnica. Un test de Vocabulario para población infantil, por ejemplo, tiene un contenido específico. El autor, luego de distintos estudios, selecciona palabras cuyo significado se le preguntará al niño/a. De todos modos, por efectos del azar, por casualidad, algún niño puede tener cierta familiaridad con una de las palabras a definir y resultarle por lo tanto más fácil que a otros. Como se verá en el próximo capítulo, la selección de los contenidos de los ítems debe ser muy cuidadosa, sin embargo, pueden verse afectados en mayor o menor grado por la incidencia del azar y constituirse en fuentes de error para las puntuaciones. Para dejar en claro estos conceptos, se destaca nuevamente que se trata de la incidencia del azar y no del sesgo de los ítems o falta de equidad entre distintos grupos de sujetos (véase apartado Sesgo y Equidad en cap. 5).

Las fuentes de error que pueden ocurrir durante la administración de la técnica son aquellas que tienen cierta influencia en cambios azarosos en la atención o motivación del sujeto examinado (desgano, ansiedad, experiencias anteriores); las variables relacionadas con las condiciones ambientales (temperatura, ventilación, ruido, iluminación) y las variables relacionadas con el examinador (su estilo, su comportamiento). Los ejemplos mencionados señalan algunas condiciones que pueden influir durante la administración, pero las situaciones son innumerables y las reacciones de los examinados frente a estas influencias pueden constituir una fuente de error en la medición de la variable en cuestión.

En síntesis, lo que se valora aquí es cómo influye "la suerte de sorteo", al decir de Hogan (2004), en las condiciones que están en juego al momento de administración.

Por otro lado, en las técnicas psicométricas, la subjetividad del evaluador no debe estar implicada en la puntuación, ya que la misma puede constituirse también en una fuente de error. Como veremos con detalle más adelante, se deben analizar muy bien las instrucciones para evaluar la técnica y la claridad de los criterios de evaluación. Cuanto menos explícitas y claras sean las pautas dadas por el autor, mayor será el margen dejado al juicio de quien puntúa y mayores las diferencias en los puntajes según quien le ha tocado en suerte al examinado. Dicho de otro modo, la falta de acuerdo entre distintos evaluadores puede generar, entonces, una variación no sistemática en las puntuaciones obtenidas a través de una técnica y por lo tanto, los resultados podrían variar según el examinador que le ha tocado en suerte al sujeto.

## 4.3 Confiabilidad de las puntuaciones

Es bien conocida la dificultad que implica tener vocablos disciplinares que se encuentran también en el lenguaje cotidiano. Este es el caso de la palabra *confiabilidad* que, si bien conocemos a partir de uso habitual o en metodología de la investigación, presenta ciertas especificaciones en el contexto de la Psicometría.

Al tratarse de un concepto clave en la Teoría Clásica de los Test (TCT), a continuación se citan varias definiciones. Se aclara que el listado no se realiza para contrastar opiniones, ni para dejar entrever polémicas entre los distintos autores, ya que todas definen la confiabilidad de modo correcto. El objetivo es que el lector se guíe por la que le resulte más clara y seleccione aquella que le resulte más comprensible.

Santisteban Requena (1990) señala que estudiar el concepto de *confiabilidad* implica analizar el grado de la determinación de la precisión con la que se realiza la medida. Es decir, el concepto de confiabilidad, que aparece en la terminología propia de la literatura psicométrica, es un concepto análogo al utilizado en otras ciencias bajo la denominación de *precisión*.

Martínez Arias (1995) llama *confiabilidad* a la tendencia de un objeto o un sujeto o una técnica a la consistencia en un conjunto de medidas de un atributo.

Cortada de Kohan (1999) afirma que la *confiabilidad* de un test se refiere a la consistencia, o mejor, a la coherencia de los puntajes obtenidos por los mismos individuos en distintas ocasiones o con diferentes conjuntos de ítems equivalentes.

Cohen y Swerdlik (2000) refieren el término *confiabilidad* a la proporción de la varianza total de las puntuaciones obtenidas con un test que puede atribuirse a la varianza verdadera. Cuanto mayor es esta proporción, más confiable es la técnica (en el siguiente apartado se define varianza verdadera y varianza de error para los lectores que necesiten revisar estos conceptos).

La definición de Anastasi y Urbina (1998), en el ya clásico texto *Tests Psicológicos*, señala que el término *confiabilidad* se refiere a la consistencia de las puntuaciones obtenidas por las mismas personas cuando se las examina en distintas ocasiones con el mismo test, con conjuntos equivalentes de ítems o en otras condiciones de administración. El concepto, agregan, fundamenta el cálculo del *error de medición* de un solo resultado, con el que podemos predecir la probable fluctuación en la calificación de un solo individuo debida a factores aleatorios irrelevantes o desconocidos.

Entre las definiciones arriba listadas, la última hace explícita la relación entre confiabilidad y error, relación que será desarrollada en el apartado 3.5., por el momento sólo se subraya el hecho de que cuando se mide un atributo psicológico –aunque esta medida sea confiable–, resultará afectada por cierta cantidad de error *aleatorio*.

Otra vía para la comprensión de este concepto es mencionar algunas cuestiones relacionadas, ya que los estudios acerca de la confiabilidad de una técnica se han desarrollado para responder interrogantes acerca de la precisión de sus puntajes. Estudiar la confiabilidad implica el desarrollo de métodos para analizar la respuesta a preguntas tales como: ¿cuánto fluctúan de un día a otro las puntuaciones de una prueba?; de modo más concreto, Ale y Agus, ¿obtendrían puntuaciones sustancialmente diferentes al ser evaluados hoy o dos meses después? O la siguiente pregunta, de distinto orden, ¿pueden tomarse como similares las puntuaciones obtenidas a través de una técnica que arroja resultados diferentes según quién la evalúe? Para contextuar esta cuestión podemos imaginar la siguiente situación: un adolescente es llevado a un Servicio de Psicopatología de un Hospital porque presenta síntomas de depresión. La psicóloga que lo admite, entre otras intervenciones, le administra un test destinado a valorar el grado de depresión presente en jóvenes, pero no completa la evaluación de la técnica. Al día siguiente otro profesional del equipo toma la historia clínica y evalúa el test. Más tarde, la psicóloga admisora vuelve a puntuar los resultados, ¿cada uno de estos profesionales, al evaluar la misma hoja de respuestas, obtendrá un puntaje diferente o similar?



Los estudios acerca de la confiabilidad se han desarrollado para responder estos y otros interrogantes sobre la precisión de los puntajes obtenidos a través de la aplicación de una técnica psicométrica. Se han desarrollado muchos recursos para analizar estas cuestiones.

Por último, es importante destacar, que la confiabilidad rara vez es una cuestión de todo o nada; hay diferentes grados de confiabilidad –la confiabilidad de una técnica psicométrica no se dirime en términos de confiable o no confiable–, hay diferentes tipos y grados de confiabilidad (Cohen & Swerdlik, 2000). Para avanzar hacia los procedimientos empíricos que permiten calcular el índice de confianza en las puntuaciones, el denominado *coeficiente de confiabilidad*, que informa el grado de precisión del instrumento es necesaria la comprensión de algunos conceptos estadísticos. El lector que tenga conocimientos sobre éstos puede saltear el siguiente apartado, mientras que sugerimos su lectura a aquel que necesite repasarlos.

## 4.4 Repaso de conceptos estadísticos relacionados

Si se entiende a la estadística como un conjunto de métodos para tomar decisiones inteligentes frente a la *incertidumbre* (Cortada de Kohan, 1994), se entiende también la necesidad de recurrir a ella en los temas que nos ocupan, confiabilidad y error de medición. El repaso que se propone incluye los siguientes conceptos estadísticos: (a) varianza y desvío estándar y (b) coeficiente de correlación, que serán de utilidad para la comprensión del resto del capítulo.

### Varianza y desvío estándar

Ambos conceptos son utilizados en el estudio de la *dispersión* de los puntajes obtenidos a través de la aplicación de una técnica a una muestra de sujetos. El *desvío estándar* y su relación con respecto a la media, se ha mencionado y definido en el capítulo 2 (véase apartado 2.3.). En cuanto a la *varianza*, es otra de las medidas de la variabilidad, que responde a la pregunta: ¿cómo están diseminadas las puntuaciones obtenidas? Concretamente, mide la dispersión de los datos con respecto a la media aritmética y queda definida como la suma de los cuadrados de las diferencias entre los valores de la variable y la media, ponderados por sus respectivas frecuencias (Pulido San Román, 1992). La varianza puede ser definida también, de modo más sencillo, como el promedio de las desviaciones al cuadrado respecto a la media del grupo. Sampieri *et. al.* (2000), a su vez, proponen definirla como la fluctuación o variabilidad promedio de un determinado valor de la población.

En otras palabras, siempre que un instrumento de medición es aplicado a un *grupo* de individuos se obtiene una distribución resultante (los individuos asumen diferentes valores en la variable). La variabilidad que encontramos en el conjunto de puntuaciones obtenidas puede expresarse como su varianza, que se simboliza como ($s^2$), siendo la raíz cuadrada de ésta el llamado desvío estándar, cuya notación es $s$.

Con frecuencia se utiliza la desviación estándar con fines descriptivos, pero en diversos procedimientos estadísticos, como los implicados en el tema que nos ocupa, es necesario partir de la *descomposición* de la varianza (aquí, la descomposición de las puntuaciones obtenidas a través de un instrumento de evaluación). Si $s^2$ representa la

varianza total de las puntuaciones obtenidas a partir de la administración de una técnica en una muestra de sujetos, $s_v^2$ representa la varianza verdadera y $s_e^2$ representa la varianza debida a error, entonces la relación de las varianzas puede expresarse como:

$$s^2 = s_v^2 + s_e^2 \qquad [2]$$

Ecuación que se lee como sigue: la varianza total en una distribución de puntuaciones obtenidas en una prueba es igual a la suma de la varianza verdadera más la varianza de error. La ventaja especial de la varianza es, tal como lo refleja la fórmula, permitir la descomposición en partes separadas que se combinan en forma aditiva para construir el total.

La medida de la confiabilidad de una técnica depende de la variabilidad de las puntuaciones que arroja, de su dispersión. El desafío, entonces, para el autor o adaptador de una técnica de evaluación psicológica es maximizar la proporción de la varianza total que es varianza verdadera, y minimizar la proporción de la varianza de error, porque de ese modo se acercará más a las puntuaciones verdaderas del atributo en cuestión. En esencia, cualquier condición que sea irrelevante para el propósito de la prueba es considerada variancia de error.

### Coeficiente de correlación

Se trata de un concepto estadístico fundamental para comprender los procedimientos empíricos a través de los cuales se valora la calidad de una técnica psicométrica. En esencia, un coeficiente de correlación (r) expresa el grado de correspondencia, o relación, o *covariación*, entre dos conjuntos de puntuaciones. Permite establecer el *grado* de asociación entre dos variables o entre una variable y un conjunto de otras variables, pero es siempre bivariada. Es una prueba estadística para analizar la relación entre dos variables siempre y cuando las mismas sean medidas en un nivel de intervalos o de razón (véase niveles de medición en apartado 1.2).

Se utiliza cuando la hipótesis a probar es correlacional, del tipo de "A mayor X, mayor Y", "A mayor X, menor Y", "Altos valores en X están asociados con altos valores en Y", "Altos valores en X se asocian con bajos valores de Y" (Sampieri et al., 2000). Se subraya que la asociación que se pretende calcular no es de tipo causal (el objetivo es evaluar la asociación entre las variables, no si una es causa de la otra, esta prueba estadística en sí no considera a una como independiente y a otra como dependiente).

Cuando la correlación entre dos variables es perfecta, el coeficiente de correlación es igual a uno ($r = 1$ o $r = -1$). Cuando no existe asociación alguna, es cero.

En resumen, el coeficiente de correlación se calcula a partir de las puntuaciones obtenidas en *una* muestra en *dos* variables. Relaciona las puntuaciones obtenidas de una variable con las puntuaciones obtenidas en otra variable, en los mismos sujetos. Si bien hay varios coeficientes de correlación, el que se utiliza con más frecuencia, por adecuarse a las características de las variables psicológicas, es el coeficiente de correlación lineal de Pearson, al que se lo expresa como $r_{xy}$.

Un ejemplo pensado desde el absurdo ha resultado útil para la comprensión de este concepto. Se trata de un investigador que sostiene una covariación entre el tamaño del pie de las personas y la magnitud del cociente intelectual (CI) de las mismas. Desconoce si a mayor tamaño del pie, mayor inteligencia o a mayor tamaño del pie menor inteligencia. Decide obtener evidencia empírica al respecto, por lo tanto, administra un test



de inteligencia a una muestra de sujetos a los cuales les pregunta también cuánto calzan. Tiene así dos mediciones de cada uno de los sujetos de la muestra y aplica el coeficiente de correlación para estudiar esta covariación que lo obsesiona. Si encontrase un valor cercano a -1, la interpretación adecuada es que mientras una variable aumenta (tamaño del pie) la otra disminuye (CI) (correlación perfecta negativa o inversa), si, por el contrario, encontrara un valor cercano a +1, esto indica que ambas covarían en el mismo sentido, -cuando aumenta una, aumenta la otra o cuando una disminuye, diminuye la otra- tienen una variación directa. Si algún día se llevara a cabo esta investigación se espera que este investigador encuentre un coeficiente de correlación igual a 0, ya que ambas variables, tamaño del pie y magnitud de CI, lo más probable es que no se hallen correlacionadas, sean independientes la una de la otra.

La siguiente tabla presenta algunos valores del coeficiente de correlación y la interpretación que puede realizarse de la relación entre las variables en estudio.

Tabla 4.1. Lectura de la interpretación de coeficientes de correlación

| Coeficiente de correlación | Interpretación |
|---|---|
| ***+1.00*** | ***Correlación positiva perfecta*** |
| + 0.90 | Correlación positiva muy fuerte |
| + 0.75 | Correlación positiva considerable |
| + 0.50 | Correlación positiva media |
| + 0.10 | Correlación positiva débil |
| ***0.0*** | ***No existe correlación alguna entre las variables*** |
| -0.10 | Correlación negativa débil |
| -0.50 | Correlación negativa media |
| -0.75 | Correlación negativa considerable |
| -0.90 | Correlación negativa muy fuerte. |
| ***-1.00*** | ***Correlación negativa perfecta**** |

* (A mayor X, menor Y" de manera proporcional. Es decir, cada vez que X aumenta una unidad, Y disminuye siempre una cantidad constante). Esto también se aplica a "a menor X, mayor Y".

Por último, es digno de destacar que las correlaciones son afectadas por la variabilidad del grupo en que fueron calculadas. Conforme disminuye la variabilidad de la muestra, también lo hace el coeficiente de correlación.

## 4.5 El coeficiente de confiabilidad

Las bondades psicométricas de un instrumento, como ya fuera dicho, vienen expresadas por su confiabilidad y por su validez. Cada una de ellas tiene una manera técnica de expresión. En el caso que nos ocupa, se trata de un número que indica en qué medida una técnica es confiable. Esa forma técnica de expresión es el *coeficiente de confiabilidad*, y la forma natural de obtenerlo es calculando la proporción (la razón o cociente) entre la varianza de la puntuación verdadera en una prueba y la varianza total. La siguiente figura muestra información similar a la que fuera presentada en la Fig. 3.1., pero se ha reemplazado el concepto de puntaje por el porcentaje de la varianza

total de las puntuaciones (el gráfico presenta un ejemplo en el que el 85% de la varianza es verdadera y el 15% es varianza de error).

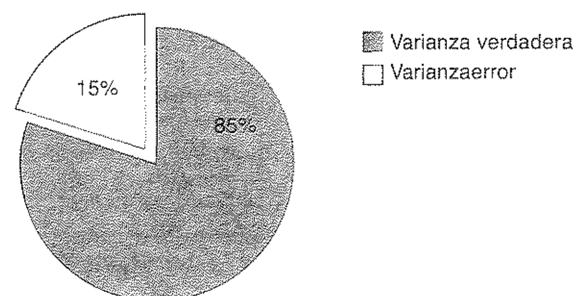



Figura 4.2. Composición de la varianza total de los puntajes

De esta forma, el coeficiente de confiabilidad -que se expresa como Cxx-, se obtiene calculando el cociente entre la varianza verdadera y la varianza total (más adelante se explica la presencia de los dos subíndices xx).

$$C_{xx} = \frac{S_v^2}{S^2} \qquad [3]$$

Nótese que si la varianza verdadera $S_v^2$ fuese igual a la varianza total $S^2$, en el diagrama 3.2 todo el círculo sería gris. En este caso, el Coeficiente de confiabilidad $C_{xx}$ sería igual a 1, que es el valor máximo que puede tomar la proporción, ya que nunca la varianza verdadera puede ser mayor que la varianza total.

Por el contrario, si la varianza verdadera $S_v^2$ fuese igual a cero, en la figura 3.2 todo el círculo sería blanco, es decir, todo sería varianza de error, entonces el coeficiente de confiabilidad $C_{xx}$ tendría un valor igual a cero, independientemente del valor de la varianza total, que, se supone, nunca es cero.

De acuerdo a esto, el coeficiente de confiabilidad es un número cuyo valor mínimo es cero –lo cual estaría indicando la inexistencia de varianza verdadera, ya que toda es varianza de error–, y su valor máximo es igual a uno –lo cual estaría indicando que no hay error, todo es varianza verdadera–.

El lector podrá comprender fácilmente que cuanto más cercano a uno sea el valor del coeficiente de confiabilidad $C_{xx}$, más confiable será el instrumento del cual se obtuvieron las puntuaciones; por el contrario, cuanto más cercano a cero es dicho coeficiente, menos confiable será el mismo.

La dificultad principal para calcular la confiabilidad, es decir, calcular esta proporción, es que el único dato que se puede obtener de los resultados de la medición corresponde a la varianza total, mientras que, tanto la varianza verdadera como la de error, son incógnitas. Por tal motivo a esta forma de calcular la confiabilidad se la denomina "forma teórica", y al coeficiente obtenido de esta manera se llama *coeficiente de confiabilidad teórico*, ya que de los tres datos de la fórmula hay dos que son desconocidos.

La existencia de una varianza verdadera se sostiene en el supuesto que ya fuera expresado: "existen *puntajes verdaderos*". Esto implica que la confiabilidad -si bien conceptualmente está bien expresada por la proporción entre ambas varianzas-,



deberá hallarse de alguna otra forma, a partir de métodos empíricos, y no con la aplicación directa de la fórmula [3]. En el siguiente apartado se revisan algunos de los procedimientos empíricos que se utilizan para obtener el coeficiente de confiabilidad.

## 4.6 Procedimientos empíricos para estimar el coeficiente de confiabilidad. Tipos de confiabilidad

Si en dos ocasiones se administra una técnica a un mismo grupo de sujetos, obteniendo de este modo dos conjuntos de medidas, el instrumento pocas veces proporcionará exactamente el mismo resultado, y esto es debido a la incidencia de factores *aleatorios*. En algunos casos las discrepancias serán grandes, en otros casos menores, pero casi siempre estarán presentes. El hecho de que las mediciones repetidas a los mismos sujetos no muestren exactamente los mismos resultados revela *falta de confiabilidad* en el instrumento. Es decir, cuanto más grandes sean las discrepancias entre la primera y segunda medición –suponiendo que el constructo no ha variado entre ambas administraciones–, menor es la confiabilidad del instrumento y mayor el error.

En consecuencia, nótese que si se realizan dos mediciones con el mismo instrumento a una muestra de sujetos (dos mediciones que pueden llevarse a cabo en forma sucesiva o simultánea), y si, además, se supone que el constructo que se quiere evaluar no varió entre las dos mediciones, entonces el conjunto de las discrepancias de los resultados entre la primera y segunda medición va a representar en alguna medida el error de medición. De este modo se cuenta con una medida de los errores, lo cual es un primer paso para hallar la varianza de error.

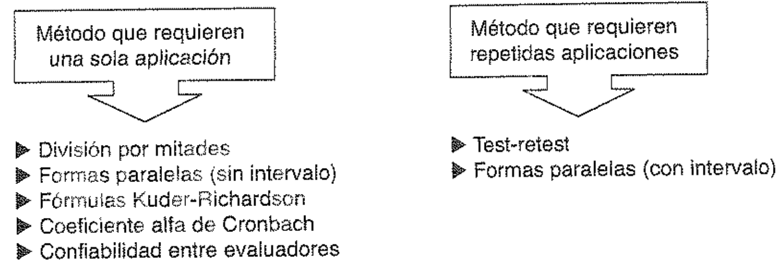
Afortunadamente, las mediciones repetidas también suelen mostrar *consistencias* (Martínez Arias, 1995), por lo cual –análogamente– las consistencias entre la primera y segunda medición llevadas a cabo con el instrumento van a representar la varianza verdadera.

En consecuencia, se han diseñado varios métodos que permiten calcular el coeficiente de confiabilidad, que, sin recurrir a la fórmula teórica, toman en consideración las discrepancias entre un primer conjunto de datos, producto de una medida, y un segundo conjunto de datos, producto de otra medición, ambas realizadas a los mismos sujetos. Dichas discrepancias guardan proporción con la varianza de error.

Existen diferentes métodos que permiten calcular empíricamente el *coeficiente de confiabilidad*, y cada uno de ellos, de acuerdo a las características del diseño, permite delimitar de manera adecuada algún aspecto sobre la confiabilidad de los puntajes arrojados por la técnica de medición en cuestión. Así, por ejemplo, si el primer conjunto de datos se toma en determinado momento y el segundo un tiempo después, y si hay muy poca discrepancia entre los conjuntos de resultados obtenidos en ambas mediciones, este dato será un indicador de que los puntajes del instrumento poseen *estabilidad* en ese lapso de tiempo.

A partir de la implementación de estos métodos empíricos, se obtienen coeficientes de confiabilidad, algunos más sensibles a la consistencia entre los ítems, otros a la estabilidad temporal del puntaje, o a la confiabilidad del evaluador, entre otras alternativas. El autor de una técnica selecciona uno u otro método dependiendo de los aspectos que le interese indagar, aunque en la actualidad es también frecuente encontrar manuales donde se informa no sólo *uno* sino varios coeficientes, o sea, se reportan *las confiabilidades* del instrumento, ya que se enfoca esta cuestión desde diversos ángulos.

El método o métodos seleccionados pueden ser clasificados de distintas maneras, en el siguiente cuadro se propone una categorización, entre otras posibles, que los divide en dos grupos, por un lado, aquellos en los que es necesaria sólo una aplicación del instrumento a la muestra de sujetos y por el otro, el grupo que requieren repetidas aplicaciones a la misma muestra.




Cada método involucra diferentes pasos para hallar el *coeficiente de confiabilidad*, pero en todos los procedimientos, como se verá más adelante, es necesario disponer de al menos dos conjuntos de *medidas paralelas de los mismos sujetos*, para luego calcular entre ellas, el coeficiente de confiabilidad de la técnica. En todos los casos se administra una técnica de evaluación psicológica a una muestra de individuos, y de este modo se obtiene una variedad de puntajes. Recordamos que aquí interesa analizar la variabilidad de las puntuaciones obtenidas por una muestra de sujetos, y no el puntaje obtenido por UN solo sujeto. El objetivo de estos estudios es analizar la *dispersión* de los puntajes que arroja la técnica administrada a un grupo de personas.

Como fuera dicho en el apartado 3.4, la correlación de Pearson es una medida que indica el grado de relación que existe entre dos conjuntos de datos, provenientes de dos variables, y su notación es $r_{xy}$. Los subíndices simbolizan las variables comparadas (véase el ejemplo de apartado 3.4., donde *x* simboliza el número de calzado, mientras que *y*, el CI.). Clásicamente se calcula la *covariación* de dos variables diferentes, sin embargo nada impide que los dos conjuntos de datos provengan de una misma variable. Por lo tanto, se puede calcular también la correlación entre un conjunto de datos provenientes de una primera medición y de una segunda, siendo la misma variable la que es evaluada en ambas ocasiones. En tal caso, el coeficiente de correlación indica la consistencia entre los puntajes obtenidos en ambas medidas: nótese que esto es justamente lo necesario para analizar la confiabilidad de las puntuaciones.

Cuanto más cercano a 1 (uno) sea el coeficiente hallado, indicará que el primer conjunto de medidas es parecido, similar, al segundo (no hubieron errores que modificaran los resultados); por el contrario, cuanto más cercano a 0 (cero) sea ese valor, indicará más discrepancia (presencia de errores) entre las dos mediciones, es decir, menos confiabilidad del instrumento (el error lo afecta en gran medida).

En consecuencia, se comprende, entonces, que la correlación entre dos conjuntos de medidas del mismo constructo es una buena medida del coeficiente de confiabilidad, ya que –a igual que este– es un número que cuando vale cero indica absoluta discrepancia entre los valores de la primera y segunda medida (máximo error), y cuando vale uno una absoluta concordancia (error nulo). Es por ello que la confiabilidad, desde el punto de vista práctico, se calcula con el coeficiente de correlación de



Pearson o –como se verá más adelante–, con variantes de éste que conservan la propiedad de valer cero cuando hay total discrepancia entre los dos conjuntos de medidas y uno cuando hay absoluta concordancia.

Cuando en el coeficiente de correlación de Pearson $r_{xy}$ se utiliza una sola variable –como en caso del cálculo de la confiabilidad–, es natural cambiar sus subíndices xy, por xx, ya que no existen dos variables, sino una variable, x, quedando entonces la notación $r_{xx}$ . Como el coeficiente de correlación hará las veces de coeficiente de confiabilidad, y, teniendo en cuenta esta nomenclatura, es que al coeficiente de confiabilidad *C* también se lo denomina con los mismos subíndices xx, expresándose $C_{xx}$. Dada la frecuencia de la utilización del cálculo de la confiabilidad a partir del coeficiente de correlación de Pearson, el lector encontrará frecuentemente, en la bibliografía, que se utilizan ambos como sinónimos (correlación de Pearson y Coeficiente de Confiabilidad), aunque en términos conceptuales no lo sean.

$$C_{xx} = r_{xx} \quad [4]$$

En razón de lo expuesto, es digna de destacar y aclarar otra cuestión, para evitar confusiones. Mientras que el coeficiente de correlación de Pearson puede asumir valores entre -1 y +1, el coeficiente de confiabilidad sólo asume valores entre 0 y +1. Es muy poco probable -y pésimo indicador de consistencia- que una correlación entre dos medidas de la misma variable en los mismos sujetos resulte con valores negativos (correlaciones inversas): en tal caso se deberá considerar que el instrumento no es confiable, que es lo mismo que decir que su confiabilidad vale cero.

Teniendo en cuenta estas aclaraciones, en la siguiente tabla se presentan los parámetros de interpretación de los datos según el coeficiente de correlación hallado, leídos en función de la confiabilidad.

Tabla 4.2. Lectura del coeficiente de confiabilidad según la correlación hallada.

| Coeficiente r de Pearson | Lectura de la correlación hallada | Lectura del coeficiente de confiabilidad |
|---|---|---|
| **+1.00** | **Correlación positiva perfecta** | Nunca alcanza este valor; ninguna técnica arroja puntajes perfectos |
| + 0.90 | Correlación positiva muy fuerte | Técnica muy confiable |
| + 0.75 | Correlación positiva considerable | Adecuada |
| + 0.50 | Correlación positiva media | Regular (no cumple requisitos científicos) |
| + 0.10 | Correlación positiva débil | Baja confiabilidad |
| **0.00** | **Ausencia de correlación entre las variables** | Medición contaminada de error. No confiable. |

Como se puede observar, se trata de una tabla similar a la 3.1. (v. apartado 3.4.), pero nótese que aquí, por un lado, se han suprimido los valores de correlación negativa, y por otro, se ha agregado la interpretación psicométrica del coeficiente de confiabilidad (y esto en tanto coeficiente aplicado al tema de la confiabilidad). La primera columna presenta un continnum que va de 0 a 1, tomando valores intermedios según su distancia a cada uno de estos polos.

Métodos basados en medidas repetidas

Es importante que un instrumento arroje mediciones estables en el tiempo. De poco serviría contar con una cinta métrica elástica, por ejemplo, que midiera los objetos de modo diferente cada vez. Obviamente, lo mismo ocurre cuando se trata de una técnica que evalúa atributos psicológicos. De poco serviría contar con un test que arrojara puntuaciones diferentes de los mismos sujetos en cada administración, y esto aún teniendo en cuenta, que los atributos psicológicos pueden ser muy dinámicos o tender a cierta estabilidad.

Una forma de estimar la confiabilidad de un instrumento de medición consiste, entonces, en usar el mismo instrumento en una muestra de sujetos, en dos momentos, es decir, con un lapso de tiempo entre ambas administraciones. Con estos métodos se estima el coeficiente de confiabilidad que permite medir la *estabilidad* de las puntuaciones obtenidas por la técnica de evaluación bajo estudio. En esta categoría encontramos el método Test-retest y el de Formas paralelas o alternativas, aplicadas con un intervalo de tiempo.

*Test-retest*

El objetivo de este método es medir la *estabilidad* de las puntuaciones sabiendo que conforme transcurre el tiempo las personas cambian. El tema de la estabilidad o cambio en las puntuaciones de una técnica psicométrica nos conduce a la siguiente cuestión, ¿se trata de un cambio real en la variable o se trata de un cambio en el puntaje debido a la falta de confiabilidad, a la inestabilidad de las puntuaciones arrojadas por el instrumento, a su sensibilidad a las fluctuaciones aleatorias?

En otras palabras, la fuente de falta de confiabilidad que identifica el método test-retest son las fluctuaciones temporales aleatorias, que influyen tanto en las condiciones de administración como en las condiciones de los examinados.

En este procedimiento empírico es fundamental la determinación de la extensión del intervalo de tiempo entre una administración y otra, ambas realizadas en una misma muestra de sujetos. La opción que tome el autor o adaptador debe estar basada en el conocimiento teórico de las características de la variable, específicamente, en el conocimiento de la evolución de la misma a través del tiempo (por ejemplo, si de la evaluación de los intereses se tratara, considerar que éstos presentan mucha inestabilidad en la niñez y tienden a estabilizarse hacia la adolescencia).

Una complicación inherente a la evaluación psicológica es que, en ocasiones, el mismo instrumento no evalúa lo mismo en dos puntos diferentes del tiempo; cuando el intervalo es breve pueden intervenir factores tales como la experiencia previa con los ítems del test, la falta de novedad, la memoria, la fatiga o la falta de motivación, y por lo tanto la segunda aplicación ya no conserva las características de la primera.

Por otro lado, por múltiples factores, conforme se incrementa el intervalo de tiempo entre las aplicaciones del mismo test, la correlación entre las puntuaciones obtenidas en cada administración tiende a disminuir.

Por lo tanto, el autor o adaptador de una técnica psicométrica, que analiza con este método la confiabilidad de la técnica en cuestión, debe explicitar claramente los criterios de selección del intervalo de tiempo entre ambas administraciones, ya que sólo así podrá ser interpretado de modo adecuado el valor del coeficiente de confiabilidad. Si el período es largo, y la variable susceptible de cambios, esto puede confundir la interpretación del coeficiente de confiabilidad obtenido por este procedimiento,



mientras que si el período es corto, las personas pueden recordar cómo contestaron en la primera oportunidad. En síntesis, cuanto mayor tiempo pase entre la primera administración y la segunda, el coeficiente de correlación será menor, y cuanto más breve sea el intervalo la estabilidad temporal de los puntajes será de menor alcance.

En el siguiente cuadro se sintetizan las etapas que se llevan adelante en la aplicación de este procedimiento.

**Etapas - Método Test-retest**

1) Aplicar y evaluar la técnica a una muestra de sujetos
2) Lapso de tiempo (justificado)
3) Aplicar y evaluar la técnica a la misma muestra de sujetos
4) Calcular la correlación (r) entre las puntuaciones obtenidas en ambas ocasiones
5) Interpretar el coeficiente hallado (estabilidad temporal de las puntuaciones)

En las tres primeras etapas se obtienen los dos conjuntos de puntuaciones de los mismos sujetos, con las cuales se calcula luego el coeficiente de correlación, el grado de asociación entre ellos. Por último, para la interpretación del coeficiente hallado se utiliza la tabla 4.2., en conocimiento de que los resultados de este método están relacionados con la estabilidad temporal de las puntuaciones.

Veamos ahora estos conceptos aplicados, concretamente, a instrumentos psicométricos. Elizabeth Koppitz (1971), en el libro *Test Guestáltico Visomotor para Niños de Bender*, presenta sus estudios sobre la confiabilidad de las puntuaciones de esta técnica. Justifica su opción por el método test-retest pero, a su vez, señala que un retest inmediato del instrumento mostraría en los resultados el efecto de la práctica en las reproducciones de los niños; mientras que un intervalo demasiado largo entre ambas administraciones reflejaría el efecto de la maduración en la capacidad visomotora en aquellos. Se decide, entonces, por un diseño donde el intervalo no sea ni muy prolongado ni muy corto y separa ambas administraciones con un intervalo de cuatro meses.

Años después, en una revisión posterior, Koppitz (1995) se refiere nuevamente a este tema, con mayor detalle y más datos. Da cuenta de los resultados reportados a partir de nueve investigaciones llevadas a cabo por distintos autores que aplican el Test de Bender con muy diferentes intervalos; el rango va desde unas horas, en el mismo día, hasta 8 meses, aunque la mayoría de los investigadores retestea a las semanas (entre 1 y 18 semanas). Las correlaciones halladas van desde 0.50 a 0.88, dependiendo del diseño. La autora concluye que estos estudios indican que las puntuaciones del Test de Bender, como técnica para evaluar la maduración perceptomotriz en niños normales escolarizados, son razonablemente estables. La confiabilidad es mayor cuando el intervalo test-retest no excede de 3 meses. Otro dato que rescata como interesante en este tema es que los niños con retrasos educativos o con una disfunción cerebral mínima tienden a madurar a un ritmo más lento y a menudo irregular. Los resultados de las investigaciones muestran concluyentemente, afirma Koppitz, que los puntajes obtenidos a través del Test de Bender, aplicado a alumnos normales escolarizados, son confiables.

En síntesis, la aplicación de este método implica una clara distinción entre la posibilidad de cambios reales en las puntuaciones de la variable, esperables desde el

punto de vista teórico y aquellos otros cambios, indicados en las puntuaciones del test, pero debidos a fuentes de error inherentes al instrumento de medición, a su falta de precisión.

Se pueden encontrar más ejemplos de aplicación de este método en los estudios realizados con el WISC-III, escala aplicable también a niños. La estabilidad de los puntajes de esta técnica fue evaluada en un estudio con intervalos que oscilaron entre 12 y 63 días, entre las dos administraciones, con un intervalo mediano de 23 días. Como muestran las tablas 5.3 a 5.5 del Manual (Wechsler, 1994), los puntajes del WISC-III muestran una adecuada estabilidad a través del tiempo y a través de los grupos de edad. En dicho texto se analizan también las discrepancias en los puntajes debidas a los efectos de la práctica entre ambas administraciones, según la duración del intervalo.

Por último, es importante señalar que la aplicación del método test-retest, para el estudio de la confiabilidad, tiene ciertas particularidades y no se lo debe confundir con los diseños en los que el método es aplicado con otros objetivos que incluyen una intervención sobre la variable durante el intervalo de tiempo entre ambas aplicaciones, tales como un proceso de aprendizaje o algún método psicoterapéutico. En estos casos se administra el test: los sujetos reciben algún tipo de entrenamiento o de tratamiento, y luego se aplica el retesteo; el diseño aquí tiene como objetivo detectar cambios en la variable, que dará cuenta a su vez de la eficacia de la intervención, por ejemplo, una mayor habilidad en la destreza aprendida o la disminución de síntomas que fueron objeto de la terapia. En este caso el método Test-retest es utilizado para captar las diferencias entre una administración y la otra. Por el contrario, en los estudios de confiabilidad, se tiene como objetivo calcular, valorar, la estabilidad temporal de las puntuaciones de la técnica, su permeabilidad a cambios sutiles y por lo tanto se espera que la intervención de factores fortuitos, aleatorios, (aprendizajes, olvidos, cambios emocionales esporádicos de los sujetos) entre la primera aplicación y la segunda influyan lo menos posible en las puntuaciones del instrumento, o sea, que el instrumento capte características constantes, estables, de los sujetos. Incluso cuando se utiliza en un diseño para captar diferencias entre un momento y otro, con una intervención específica durante ese lapso, resulta sumamente útil contar con estudios que detallen las diferencias entre la primera y la segunda administración de la técnica **sin** que intervenga un proceso, tal como fuera informado en algunas técnicas (v. Manual del WISC-III) para poder efectivamente valorar cuánto del cambio observado puede ser atribuido a la intervención y en cuánto es esperable un cambio por tratarse, meramente, de una segunda administración.

### *Formas paralelas o alternativas (con intervalo)*

Como fuera mencionado en el apartado anterior, la evaluación de la variable no conserva las mismas características cuando un test es administrado en una segunda oportunidad, ya que las respuestas a algunos ítems pueden verse afectadas por factores tales como la experiencia previa con los reactivos del instrumento, la falta de novedad, la memoria, la fatiga o la falta de motivación.

El procedimiento de las formas paralelas con intervalo de tiempo es utilizado cuando se necesita minimizar el efecto de la memoria del contenido de otra prueba aplicada con anterioridad. Es decir, es una buena alternativa cuando no se puede aplicar el método test-retest por el efecto que el aprendizaje y la memoria tendrían sobre los resultados en la segunda administración. Se procede entonces a elaborar formas equivalentes y se las aplica a los mismos sujetos en dos oportunidades, con un intervalo de



tiempo entre ambas administraciones. En este método es tan importante, como el de test-retest, justificar el lapso de tiempo, la magnitud del intervalo.

Al aplicar este procedimiento empírico no se administra el mismo instrumento de medición, sino dos formas equivalentes, una en cada sesión. El autor del test debe elaborar una técnica y otra equivalente, su forma paralela. Su tarea es similar a la de un profesor que prepara los temas A y B de un examen parcial. Ambas versiones deben partir de un fundamento común, –la bibliografía del programa–, tener un contenido y un grado de dificultad similar, sin ser iguales.

Del mismo modo, las formas paralelas de una técnica deben ser similares en contenido, instrucciones y duración, pero también deben ser equivalentes, tanto en las medias y las varianzas de las puntuaciones que arrojan, como en los índices de dificultad y discriminación de los ítems.

En síntesis, este procedimiento controla dos fuentes de falta de confiabilidad, las fluctuaciones temporales aleatorias (al igual que el método anterior, el test-retest) y además la inconsistencia de las respuestas a diferentes muestras de ítems, ya que hay cambios en los reactivos del instrumento administrado en la primera sesión y en la segunda sesión. El conjunto de ítems que conforman una técnica psicométrica es una muestra seleccionada de todos los reactivos posibles; en el caso de las formas paralelas, el autor elabora dos muestras de ítems similares, que, al ser aplicadas pueden ser más o menos consistentes entre sí.

En cuanto a las etapas del procedimiento, son similares a las del método test-retest, aunque en su interpretación se agregan cuestiones relacionadas con el contenido de ambas formas de la prueba.

**Etapas - Formas Paralelas (con intervalo)**

1) Administrar una forma del test a una muestra de sujetos
2) Lapso de tiempo (justificado)
3) Administrar la forma paralela del test a los mismos sujetos
4) Calcular la correlación (r) entre las puntuaciones obtenidas con una forma y con la otra
5) Interpretar el coeficiente hallado (estabilidad temporal de las puntuaciones y muestreo de contenido)

Al interpretar el coeficiente de correlación hallado se debe de tener en cuenta la influencia tanto de la estabilidad temporal de los puntajes (por el intervalo de tiempo entre una aplicación y la otra) como el muestreo de contenido (debido a la probable influencia de razones azarosas en la selección de ítems que componen una forma y la otra).

## Métodos basados en una sola aplicación del test

Los métodos que implican la administración de la técnica en una muestra en un momento determinado, son los más utilizados por los autores y adaptadores de las técnicas psicométricas.

En los siguientes apartados se analizan cinco métodos: división por mitades; formas paralelas o alternativas (sin intervalo de tiempo); las fórmulas Kuder Richardson; el coeficiente alfa de Cronbach y la confiabilidad entre evaluadores.

*División por mitades*

El autor o adaptador de una técnica psicométrica que utiliza este método, tiene como objetivos el escrutinio de los ítems que conforman la prueba y el análisis de las relaciones entre ellos. El procedimiento empírico aporta información para estimar el grado de *consistencia interna* del instrumento. En otras palabras, el método división por mitades controla o identifica la inconsistencia de la muestra de ítems, el muestreo de contenido. Es condición que la técnica en estudio sea homogénea, que evalúe un único atributo o factor.

Requiere sólo una aplicación del test a una muestra de sujetos, luego de lo cual se procede a dividir la prueba en mitades homogéneas, apareadas en contenido y dificultad. Pero esta partición de los ítems no es una mera división de los reactivos en dos mitades. Existen diversas formas adecuadas para logar dos mitades homogéneas. Una forma aceptable es asignar cada ítem, a una mitad o a la otra, al azar. Otra modalidad, muy utilizada en tests de aptitudes, consiste en dividirlos en números pares e impares, de modo que los ítems quedan ordenados según su dificultad creciente, ya que de no usar este criterio, al segmentar un instrumento de 20 reactivos en dos mitades formadas por los primeros 10 y otra integrada por los últimos 10, quedaría una mitad del test con los ítems de baja dificultad y la otra, sólo con los difíciles. Otra alternativa es dividir la prueba por contenidos, de modo que cada mitad del test contenga ítems equivalentes en cuanto al contenido y la dificultad.

A su vez, las mitades deben ser similares en cuanto a formato, número de ítems y estadísticos (medias, varianzas e índices de dificultad y discriminación), en síntesis, deben ser homogéneas.

El siguiente cuadro sintetiza las etapas involucradas en este procedimiento.

**Etapas - División por Mitades**

1) Aplicar la técnica a una muestra de sujetos
2) Dividir el conjunto de ítems en dos mitades homogéneas
3) Calcular la correlación (*r*) entre las puntuaciones obtenidas en las dos mitades en las que ha quedado dividida la técnica
4) Ajustar la confiabilidad de la mitad de la prueba usando la fórmula de Spearman-Brown
5) Interpretar el coeficiente hallado (consistencia de las respuestas a lo largo del test)

En la tercera etapa se calcula la correlación (covariación) entre los puntajes de cada una de las mitades que han sido aplicadas a la misma muestra.

La cuarta consiste en la aplicación específica de la fórmula Spearman-Brown que se utiliza para estimar la confiabilidad de un instrumento cuando éste se ha alargado o acortado en cualquier cantidad de ítems. En este caso, la división ha acortado los ítems a la mitad, entonces, el cálculo de la confiabilidad de la consistencia interna debe ser ajustado al test entero (véase con mayor detalle las aplicaciones de la fórmula Spearman-Brown en Cohen y Swerdlik, 2000).

Veamos, a partir de algunos ejemplos, cómo el autor de una técnica valora la posibilidad de utilizar o no este procedimiento. Koppitz (1971), en el libro *Test Guestaltico Visomotor para Niños* descarta el método de división por mitades. Este procedimiento, aclara, no es apropiado para verificar la confiabilidad del instrumento en cuestión, sencillamente porque es imposible dividir en dos mitades homogéneas las 9 tarjetas que conforman esta prueba.



Por el contrario, en el estudio de las propiedades de algunos subtests del WISC-III (Wechsler, 1994), este procedimiento resulta adecuado ya que los distintos subtests pueden ser divididos en dos mitades homogéneas.

¿Por qué se estudiaron con este método sólo "algunos subtests" y no el WISC completo? Esta cuestión permitirá aclarar una condición para aplicar este procedimiento, a la que hemos aludido al comienzo de este apartado. El WISC es un instrumento heterogéneo, mide diversos atributos y factores, por lo tanto sería imposible dividirlo en dos mitades homogéneas. La *homogeneidad*, –palabra derivada del griego *homos*, que significa "misma" y *genous*, que significa "clase"–, indica, en términos psicométricos, que una técnica mide un solo atributo o factor.

Al interior de cada test sí es posible aplicar la división por mitades en "algunos subtests", podríamos preguntar también ¿por qué no aplicarlo en todos? Porque dos de ellos –Claves y Búsqueda de Símbolos– son pruebas de velocidad. En este tipo de pruebas, las diferencias individuales en las puntuaciones dependen de modo fundamental de la velocidad en la ejecución, los ítems son muy fáciles, siendo, además, el puntaje total igual al número de reactivos que el sujeto logró realizar hasta el tiempo límite. Ya que el contenido no es relevante –se trata de ítems muy fáciles– y es similar a lo largo de toda la prueba, el muestreo del contenido incluido en cualquiera de las dos mitades es igual, luego la correlación entre la mitad de los ítems pares y la mitad de ítems impares va a ser perfecta (+1), pero al mismo tiempo espuria, ya que no aportaría información sobre la confiabilidad de las puntuaciones. En consecuencia, este procedimiento y otros similares a él, es inapropiado para valorar la confiabilidad de las pruebas de velocidad.

Hechas estas aclaraciones y volviendo al WISC, la confiabilidad de la mayoría de los subtests –excepto Claves y Búsqueda de Símbolos– fue calculada con este método. Los ítems de los subtests fueron divididos en dos medios tests con variancias aproximadamente iguales, luego se correlacionaron los puntajes de los dos medios tests, y el coeficiente resultante fue corregido por la fórmula Spearman-Brown. En el Manual se reportan los coeficientes de confiabilidad corregidos: Información, 0,84; Analogías, 0,81 y Vocabulario, 0,87. Todos los coeficientes de confiabilidad informados surgen del promedio de los coeficientes hallados en estos subtests en las distintas edades (v. Tabla 5.1. en pág. 202 del Manual de la Tercer versión del Test de Inteligencia para niños de Wechsler, 1994), ya que la segmentación de la muestra de sujetos por edad dio lugar a la formación de distintos grupos y el coeficiente confiabilidad se halla sujeto a la variabilidad de cada grupo.

*Formas paralelas o alternativas (sin intervalo)*

Este método es similar al de Formas paralelas o alternativas, con intervalo de tiempo, la diferencia reside en que en este procedimiento no hay un lapso de tiempo que separe ambas administraciones. Se aplican ambas formas –que, ya se ha explicado, deben ser equivalentes–, en la misma sesión, a la misma muestra de sujetos, una después de la otra.

Este procedimiento controla específicamente si razones azarosas en la selección de los ítems de cada una de las formas han influido en la muestra de sujetos de tal manera que los mismos contestan mejor en una forma específica del test que en la otra, y esto, obviamente, no en función de variaciones verdaderas en el constructo a evaluar, sino

tan sólo debido a que los ítems particulares que le tocaron en suerte, o sea, por el azar, por la influencia de errores aleatorios. En síntesis, identifica la presencia de inconsistencias en las respuestas a diferentes muestras de ítems. El procedimiento es el que sigue.

**Etapas - Formas Paralelas o alternativas (sin intervalo)**

1) Aplicar las dos formas a una muestra de sujetos (sin intervalo de tiempo entre ambas).
2) Calcular la correlación (r) entre las puntuaciones obtenidas por la misma muestra en una y otra forma.
3) Interpretar el coeficiente hallado (consistencia de las puntuaciones).

Se debe de tener en cuenta que las aplicaciones de ambas formas, sin intervalo de tiempo, pueden ser afectadas por la fatiga y/o la falta de motivación por parte de los sujetos (las sesiones de administración suelen ser de larga duración). Por otro lado se deben contemplar las diferencias que pueden deberse al orden de aparición de cada una de la formas en la aplicación, es decir, si la forma A o la B ha ocupado la primera posición o la segunda.

*Fórmulas Kuder-Richardson*

La insatisfacción con los métodos de división por mitades llevó a Kuder y Richardson a desarrollar sus propias medidas para estimar la confiabilidad. Se trata de índices útiles para evaluar la *homogeneidad* del test. Estas fórmulas permiten calcular el grado de correlación entre todos los ítems de una escala. Mencionamos aquí el *Coeficiente KR-20* (llamado así debido a que es la vigésima fórmula desarrollada en una serie). Se trata de variantes del coeficiente de correlación de Pearson, para ser utilizados en casos especiales.

Cuando los ítems de un test son muy homogéneos, las estimaciones de confiabilidad KR-20 y de división por mitades serán similares. Sin embargo, la KR-20 es la estadística seleccionada cuando se desea determinar la consistencia entre ítems dicotómicos, sobre todo aquellos ítems que pueden ser calificados como correctos o incorrectos. En estos casos, este método identifica la inconsistencia entre los ítems, la cual puede estar influida por el muestreo de contenido o por la heterogeneidad del atributo evaluado. Sus etapas son las que siguen.

**Etapas – Método Kuder Richardson**

1) Aplicar y evaluar la técnica a una muestra de sujetos.
2) Calcular el coeficiente KR-20 entre los ítems.
3) Interpretar el coeficiente hallado (consistencia, homogeneidad).

La fórmula K-R20 tiende a ser anticuada en una época de programas específicos de computación. Aunque se han propuesto numerosas modificaciones a esta fórmula, la que ha recibido la mayor aceptación hasta la fecha es un estadístico llamado *coeficiente alfa*, al que le dedicaremos el siguiente apartado.



*Coeficiente alfa de Cronbach*

Mientras que la fórmula *K-R20* se usa en forma apropiada con ítems dicotómicos, el *coeficiente de Cronbach* puede ser utilizado en reactivos no dicotómicos, o sea, en ítems que incluyen un rango de alternativas posibles para que el sujeto los responda (por ejemplo, las escalas Likert que se mencionan en el cap. 1), y en los que, además, suelen incluir créditos parciales.

El *coeficiente alfa*, desarrollado por Cronbach en 1951, ampliado por Novick y Lewis en 1967 y por Kaiser y Michael en 1975, resulta muy ventajoso y es ampliamente utilizado. En la actualidad, es el estadístico preferido para obtener una estimación de la confiabilidad de la consistencia interna. Puede considerarse como la media de todas las correlaciones de división por mitades posibles, cumplan o no con los requisitos del método de división por mitades, que luego serán corregidas por la fórmula de Spearman-Brown. El procedimiento es sencillo y el cálculo muy fácil si el investigador es asistido por un programa informático adecuado. Se trata de un método para identificar inconsistencia entre los ítems de una técnica.

**Etapas - Coeficiente alfa de Cronbach**

1) Aplicar la técnica a una muestra de sujetos.
2) Calcular el coeficiente alfa ($\alpha$) entre las puntuaciones obtenidas en los distintos ítems.
3) Interpretar el coeficiente hallado.

Un ejemplo: Millon (1997) informa la confiabilidad del *Inventario de Estilos de Personalidad* en sujetos adultos establecida mediante el coeficiente alfa promedio. Informa un resultado igual a 0,775 en la escala Innovación. En la actualidad, el coeficiente final debe ser acompañado por el rango de los coeficientes parciales, así, en el MIPS, arrojó el valor mínimo igual a 0,69 y el valor máximo igual a 0,85, en la escala analizada.

*Confiabilidad entre evaluadores*

Una técnica psicométrica confiable debe arrojar los mismos resultados independientemente de quien lleve a cabo la evaluación, ya que la medición es estandarizada e implica uniformidad tanto en las condiciones de administración como en las de evaluación. Numerosas técnicas se puntúan de manera directa y objetiva, pero, en ocasiones, el autor debe analizar si la técnica cumple con este requisito de "objetividad" (o de la incidencia mínima de la subjetividad del evaluador). Este método identifica las fluctuaciones en las puntuaciones según el evaluador. Sus etapas son las que siguen.

**Etapas - Confiabilidad entre evaluadores**

1) Administrar la técnica a una muestra de sujetos.
2) Evaluar las técnicas administradas (Evaluador A)
3) Evaluar las técnicas administradas (Evaluador B)
4) Calcular la correlación (r) entre los puntajes asignados por Evaluador A y por Evaluador B.
5) Interpretar el coeficiente hallado

Por ejemplo, para puntuar el subtest de Vocabulario (listado de palabras que el niño tiene que definir), el Manual del WISC-III provee ejemplos de respuestas correctas, parcialmente correctas e incorrectas para evaluar las respuestas del sujeto, pero es prácticamente imposible que pueda dar cuenta de todas las respuestas posibles. Por lo tanto, fue necesario estudiar la confiabilidad interexaminadores. En este caso, entonces, se administró el subtest Vocabulario a una muestra de niños. Luego los resultados fueron evaluados por un mínimo de dos examinadores, hasta un máximo de cuatro. Se calculó la correlación entre los puntajes por ambos –o interclase cuando fueron cuatro– y se halló un coeficiente igual a 0,98. Este resultado permite afirmar que este subtest puede ser puntuado con elevada confiabilidad.

Resumiendo, con este método se analiza la concordancia entre los evaluadores, y el autor o adaptador de una técnica psicométrica lo selecciona cuando la puntuación de los ítems del test en cuestión es compleja y/o requiere cierto grado de elaboración por parte del examinador.

Este coeficiente de confiabilidad informa al usuario de la técnica que las puntuaciones pueden derivarse en forma consistente y sistemática cuando distintos evaluadores siguen las instrucciones dadas en el Manual.

Cuando el elaborador o adaptador de una técnica calcula un índice bajo de confiabilidad con este método deberá revisar los criterios de puntuación e incluir otros que resulten más claros y que permitan, por lo tanto, obtener un coeficiente mayor.

Por último, y a modo de síntesis, en este apartado nos hemos preguntado, ¿cuál es la *utilidad* del coeficiente de confiabilidad? Estamos en condiciones de responder que es útil para conocer ciertas propiedades psicométricas de una técnica, para valorarla, para tener criterios de selección entre instrumentos. En el siguiente veremos que es útil también para calcular el error de medición de las puntuaciones obtenidas a través una técnica e interpretarlas adecuadamente.

## 4.7 Error típico de medida. Su utilidad

El coeficiente de confiabilidad ayuda al autor o adaptador de una técnica a construir un instrumento de medición adecuado, mientras que al administrador o usuario lo ayuda a valorar cuándo una técnica es adecuada para evaluar una variable de su interés. Sin embargo, la utilidad de este índice no termina en la construcción y la selección de la técnica (Cohen y Swerdlik, 2000). El administrador de un test debe conocer el error que comente al realizar una medida, y es el coeficiente de confiabilidad el que permite su cálculo.

En otros términos, en este apartado veremos cómo la confiabilidad es importante a la hora de interpretar *puntuaciones individuales*. En este tema ya no se analiza la variabilidad de las puntuaciones obtenidas en una muestra de sujetos para estudiar el instrumento, sino que se aborda el análisis de una puntuación específica de un sujeto concreto. Se brindan, además, herramientas útiles para todo aquel que incluya técnicas de evaluación psicológica en sus actividades profesionales y para aquellos que necesiten leer información obtenida a través de técnicas psicométricas.

Como ya hemos dicho, en la práctica es muy poco frecuente que una medición sea perfecta. Generalmente tiene algún grado de error. Desde luego, se trata de que este error sea el mínimo posible, ya que cuanto mayor sea el error al medir, el valor obtenido a través de la técnica se alejará más del valor real o verdadero.



Recordemos aquí que la puntuación *verdadera* de un sujeto rara vez puede determinarse exactamente, lo más probable es que la misma pueda ser estimada a partir de las puntuaciones observadas.

En sentido amplio, el error se refiere al componente de la puntuación obtenida por un sujeto en una técnica psicométrica, que no está en relación con la evaluación del atributo en cuestión (por ejemplo, en un test de inteligencia, parte del puntaje obtenido por un sujeto tiene un componente que no evalúa su inteligencia, sino otras variables improcedentes). Más aún, no se puede saber si el puntaje obtenido subestima o sobreestima el atributo evaluado. Es decir, el error se refiere al componente de la puntuación observada que está evaluando dichas variables improcedentes, condiciones aleatorias y no permanentes del atributo en cuestión.

En el apartado 4 estos conceptos se han formalizado en las expresiones

$$\mathbf{X_v = X - X_e} \qquad [1]$$

Donde sólo se cuenta con información sobre el valor medido **X**, ya que tanto el componente verdadero $\mathbf{X_v}$ como el de error $\mathbf{X_e}$, son desconocidos.

También se presentó la siguiente fórmula

$$\mathbf{s^2 = s_v{}^2 + s_e{}^2} \qquad [2]$$

Expresión en la que ocurre algo análogo que en [1], es decir, se puede acceder al cálculo de la varianza total de las puntuaciones, pero no a los valores de sus componentes de varianza verdadera y de error.

Por otra parte, se ha dicho que la confiabilidad se puede definir como la proporción de la varianza verdadera y la total.

$$C_{xx} = \frac{S_v^2}{S^2} \qquad [3]$$

despejando de [2] $S_v^2 = S^2 - S_e^2$ y reemplazándolo en [3]

$$C_{xx} = \frac{S_v^2 - S_e^2}{S^2}$$

lo que puede escribirse como...

$$\boxed{C_{xx} = 1 - \frac{S_e^2}{S^2}} \qquad [5]$$

En esta última expresión, la confiabilidad es un dato que puede ser calculado a partir de algunos de los procedimientos empíricos descriptos, mientras que la varianza de los puntajes, como se indicó, es un dato que puede calcularse. Es decir, aquí hay solo una incógnita que es $S_e$; entonces, despejando de la anterior...

$$S_e = S\sqrt{1 - C_{xx}} \quad [6]$$

El símbolo $S$, desviación típica o estándar, representa la variabilidad, en este caso, de los puntajes obtenidos cuando se aplica el test a un conjunto de individuos. Por otro lado, como puede observase, cuanto mayor sea el coeficiente de confiabilidad, menor será el error típico de medición, ya que a medida que aumenta $C_{xx}$ –el coeficiente de confiabilidad–, el segundo término del segundo miembro de la igualdad disminuye. Si llegara $C_{xx}$ a ser igual a 1, entonces el $s_e$ –el desvío estándar del error de medición–, sería igual a cero. Por el contrario, si la confiabilidad fuera $C_{xx}$= 0, entonces el desvío estándar del error sería igual al de la variable medida, es decir, todo sería error.

Con esta expresión se puede calcular el desvío típico del error, que, como fuera señalado, en el repaso de los conceptos estadísticos, es un valor que indica el promedio de la dispersión de los puntajes –en este caso de los errores– alrededor de su valor promedio. Es muy importante recordar aquí que el promedio de los puntajes de error vale cero, es decir en un número elevado de mediciones se producirán tantos errores por exceso como por defecto, dando el promedio cero.

En el capítulo 2, ya se ha hecho referencia al hecho de que gran cantidad de variables se distribuyen de acuerdo a la curva normal, pues el error de las mediciones es una de ellas.

Sabiendo entonces que el error tiene una distribución normal, cuyo valor es cero y su desvío es calculable, y conociendo, además, la confiabilidad de la técnica, solo resta hacer un breve repaso de la curva normal para poder aprovechar sus propiedades.

En el apartado 4.4 fue dicho que entre un desvío por encima de la media y uno por debajo se encuentran el 68 % de los puntajes, y que entre dos desvíos estándar por encima de la media y dos por debajo de la misma se encuentran el 95 % de los puntajes; ya entre tres desvíos por encima y por debajo de la media se encuentran el 99 % de los puntajes. Los intervalos así delimitados indican la posibilidad de que un determinado puntaje se encuentre dentro de ellos. Así por ejemplo, hay un 68 % de posibilidades de que cometido un error este se encuentre entre más o menos un desvío estándar de la media de error (cero), y hay un 95 % de que ese error esté entre dos desvíos estándar por encima y debajo de la media, y así consecutivamente.

## Niveles de significación e intervalo de confianza

Ya se ha destacado que no es posible calcular el error de una determinada medición (ya que no se conoce el valor verdadero), sin embargo, calculando el desvío estándar de los errores se puede conocer la posibilidad de que el error se encuentre entre dos determinados –y calculables– valores. A estos dos valores- uno por encima del puntaje obtenido y otro por debajo del mismo-, con su correspondiente probabilidad, se los conoce como *intervalo de confianza*. Así, por ejemplo, obtenido un determinado puntaje producto de una medición, se puede asegurar con el 68 % de certidumbre que el puntaje verdadero estaría entre un desvió estándar de error por encima y uno por debajo de dicho puntaje.

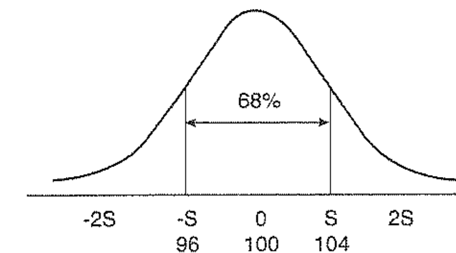


Veamos estos conceptos aplicados a un ejemplo. Supongamos que a un niño, al que le dicen Pipo, se le administró un test que evalúa la Inteligencia Musical de las personas y obtuvo un puntaje de 100 puntos. El test en cuestión presenta un desvío típico $S$ igual a 10, siendo el coeficiente de confiabilidad $C_{xx}$ igual a 0,84. Se aplica la fórmula [6], y se obtiene el valor del desvío estándar del error de medición del instrumento.

$$s_e = 10\sqrt{1 - 0{,}84} = 4$$

Entonces, si el resultado de la medición (valor medido, puntaje obtenido) ha sido de 100 puntos, se puede asegurar, con un 68 % de certeza, que el valor verdadero estaría entre los valores 96 y 104 puntos, ya que éstos toman en cuenta un desvío estándar de error por encima y por debajo del puntaje obtenido.

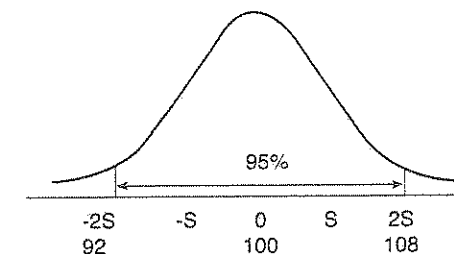
Esto podía graficarse de la siguiente manera:




*En el gráfico pueden observarse el valor medio de 100 puntos, un desvío estándar de error por debajo (96) y uno por encima (104): el área de la curva entre ambos valores, representa la probabilidad de que el valor verdadero se encuentre entre esos valores: en la curva normal, ese valor es del 68%.*

De la misma forma, podría indicarse que si el resultado de la medición (valor medido, puntaje obtenido) fue de 100 puntos, se puede asegurar con un 95 % de certeza, que el valor verdadero estaría entre los valores 92 y 108, que señalan los dos desvíos estándar de error por encima y por debajo del valor medido.

Eso podría graficarse de la siguiente manera:




*En el gráfico pueden observarse el valor medio de 100 puntos, dos desvíos estándar de error por debajo (92) y dos por encima (108): el área de la curva entre ambos valores, representa la probabilidad de que el valor verdadero se encuentre entre ellos: en la curva normal, ese valor es del 95%.*

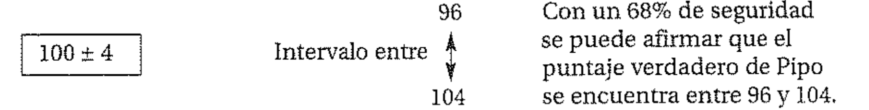
Como se puede observar, si bien el desvío estándar del error no permite precisar cuál es el error que se comete en una determinada medición, permite sin embargo calcular los valores de los intervalos de confianza, es decir, estimar con una determinada probabilidad entre qué puntajes estaría el valor verdadero. En vista de esta utilidad, al desvío estándar de error se lo denomina como ***error estándar*** o ***error típico***. Dicho de otro modo, el error típico así calculado (igual a 4 en el ejemplo), permite estimar el rango de la puntuación verdadera, o sea, las puntuaciones entre las cuales se encontrará, con cierto grado de probabilidad, el puntaje verdadero del sujeto.

Estos conceptos pueden resumirse en el siguiente cuadro.

1) Se espera que aproximadamente el 68% de las puntuaciones ocurran dentro del intervalo dentro de ± 1$\sigma$*med*.
2) Se espera que aproximadamente el 95% de las puntuaciones ocurran dentro del intervalo dentro de ± 2$\sigma$*med*.
3) Se espera que aproximadamente el 99% de las puntuaciones ocurran dentro del intervalo dentro de ± 3$\sigma$*med*.

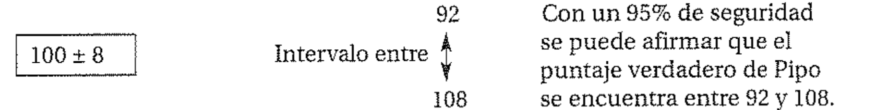
Si se sigue este esquema en el ejemplo dado, el puntaje obtenido por Pipo, sería:

1) Se desea tener una seguridad del **68%** de que el puntaje verdadero de Pipo se encuentre en el intervalo de puntajes hallado. Entonces se suma y se resta un error típico de medición.



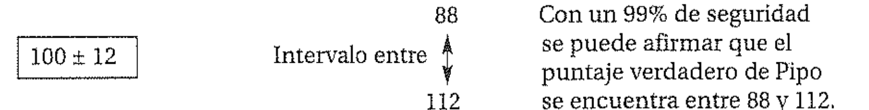

Con un 68% de seguridad se puede afirmar que el puntaje verdadero de Pipo se encuentra entre 96 y 104.

2) Se desea tener una seguridad del **95%** de que el puntaje verdadero de Pipo se encuentre en el intervalo de puntajes hallado, entonces se suman y se restan 2 errores típicos de medición.




Con un 95% de seguridad se puede afirmar que el puntaje verdadero de Pipo se encuentra entre 92 y 108.

3) Se desea tener una seguridad del **99%** de que el puntaje verdadero de Pipo se encuentre en el intervalo de puntajes hallado. Entonces se suman y se restan 3 errores típicos de medición.




Con un 99% de seguridad se puede afirmar que el puntaje verdadero de Pipo se encuentra entre 88 y 112.



Como el lector habrá podido observar, a medida que aumentamos la seguridad, la certeza, la confianza en la evaluación, aumenta también el rango del intervalo (la distancia entre los puntajes mínimo y máximo del intervalo).

## Utilidad del error típico de medida

Para facilitar el cálculo del error estándar y los intervalos de confianza que a partir de él pueden obtenerse, algunos manuales de las técnicas de evaluación psicológica, proveen estos valores tabulados de acuerdo a algunos niveles de confianza de uso práctico en psicología. El WISC-III (Wechsler, 1994) es un ejemplo.

Esta técnica estima el rendimiento general de un niño o adolescente a través del Cociente Intelectual de la Escala Completa (CIEC), Verbal (CIV), de Ejecución (CIE) y cuatro puntajes Índice [Comprensión Verbal (CV), Organización Perceptual (OP), Ausencia de Distractibilidad (AD) y Velocidad de Procesamiento (VP)]. Las interpretaciones, tanto cuantitativas como cualitativas, de los puntajes específicos, deben tener en cuenta el error de estimación inherente a los datos obtenidos a través de este test. De hecho, el Manual proporciona los intervalos de confianza para que el usuario pueda estimar la precisión de los puntajes y, por lo tanto, conozca la gama de valores en la que probablemente se encuentra el verdadero puntaje del niño o adolescente evaluado. En diversas tablas, el manual ofrece los intervalos de confianza a dos niveles de significación, 0,90 y 0,95. En otras palabras, con un porcentaje de certeza del 90% y del 95%, respectivamente.

A modo de ejemplo, en la siguiente tabla se han convertido los puntajes directos de un niño de 9 años, 2 meses, cuyo supuesto nombre es Agus.

Tabla 4.3. Ejemplos de intervalos de confianza. Fuente: Wechsler, 1994

| | Puntaje Bruto | CI | Intervalo de confianza 90% | 95% |
|---|---|---|---|---|
| Verbal | 59 | 111 | 105-116 | 104-117 |
| Ejecución | 44 | 93 | 87-100 | 86-102 |
| Completa | 103 | 102 | 97-107 | 96-108 |
| C.V. | 45 | 107 | 101-112 | 100-113 |
| O.P. | 38 | 97 | 90-104 | 89-106 |
| A.D. | 29 | 126 | 115-130 | 113-132 |
| V.P. | 23 | 109 | 100-116 | 98-117 |

En la tercera columna se encuentran los CI correspondientes a los puntajes brutos, mientras que la cuarta y la quinta proveen información sobre los intervalos de confianza a nivel del 90% y del 95%, respectivamente, de los Cocientes Intelectuales (en el caso del Manual del WISC informa directamente la banda de puntuaciones que con alta probabilidad contiene la puntuación verdadera).

Nótese en la tabla, en principio, que, a diferencia de lo explicado en el apartado anterior, no se encuentra el nivel de certeza del 68%. Esto es así ya que este nivel de certidumbre es muy bajo y, por lo general, no es utilizado en mediciones científicas. Con respecto a otros niveles de medición a los que nos hemos referido, está presente

el de un 95% de certeza y no figura el del 99%. Esto se debe al hecho de que un nivel de certeza del 99% es muy elevado, teniendo en cuenta el desarrollo actual de las técnicas psicométricas. Por lo tanto, el WISC asume dos niveles de significación de sus puntuaciones, una que implica un nivel de certeza del 90% y otra del 95%.

Con respecto a la lectura de los datos del ejemplo que se viene ilustrando, el CI Verbal obtenido por Agus es igual a 111, por lo tanto se puede tener un 90% de confianza en que el verdadero puntaje del CIV se encuentra en la franja comprendida entre 105 y 116. Si se aumenta a 95% el porcentaje de confianza, el intervalo es más amplio aún; ya que es altamente probable que el verdadero puntaje se encuentre entre los valores que van de 104 a 117).

Es frecuente que un extremo del intervalo tenga una interpretación diagnóstica (CIV 105 = Inteligencia Verbal promedio) y el otro una diferente (CIV 116 = Inteligencia Verbal media alta), al quedar incluido en la franja el puntaje de corte (CI = 110) entre los diagnósticos "Promedio" y "Rendimiento Medio Alto".

En los resultados obtenidos por Agus, se presenta esta situación en la interpretación diagnóstica tanto del CIV como del CIE; no así en el CIEC en la que ambos extremos del intervalo (97-107) remiten a la descripción cualitativa "Inteligencia Promedio".

Este ejemplo permite valorar la importancia de no regirse por la lectura puntual del puntaje obtenido en una técnica ya que el margen de error puede confundir un diagnóstico. En este caso, como en otros, el análisis cualitativo de las respuestas del niño y su rendimiento en el resto de la prueba, junto con otros datos, tales como su historia de vida o su contexto, serán decisivos para que el profesional llegue al diagnóstico pertinente.

La evaluación psicológica en numerosas ocasiones está ligada a la toma de decisiones y cada vez es más frecuente la solicitud de los intervalos de confianza en los informes psicológicos. Ejemplo de ello son los informes solicitados en ámbitos escolares, por ejemplo, ante el diagnóstico de retraso mental o talento, así como en los peritajes judiciales.

Antes de abandonar el ejemplo presentado en la Tabla 4.3, queremos hacer notar que los intervalos de confianza son bastante simétricos, sin serlo a la manera en que se ha explicado en el apartado anterior. En ocasiones no lo son, y para aclarar esta cuestión ahora pondremos el foco en otro ejemplo, el de un niño que ha obtenido un CI igual a 70 en la Escala Completa. En este caso, la tabla que encontraríamos sería la siguiente:

Tabla 4.4. Ejemplo de intervalos de confianza asimétricos

| | CI | Intervalo 90% | Intervalo 95% |
|---|---|---|---|
| Escala completa | 70 | 66 – 76 | 66 - 77 |

Como puede observarse, el valor mínimo del intervalo de confianza a un nivel de confianza del 90%, es igual a 66. Se encuentra, entonces, 4 puntos por debajo del puntaje obtenido (CI= 70), mientras que el valor máximo del intervalo, igual a 76, se encuentra a 6 puntos por encima de éste.

Que el error sumado y restado al puntaje obtenido no sea el mismo se debe a que existen diferentes metodologías para el tratamiento del error de medición. En el caso del WISC, a partir de su tercera versión, los intervalos de confianza fueron elaborados



con una metodología ligeramente diferente, que se basa en el error típico de *estimación*, y no en el error típico de medición. El error en la que se basa esta técnica se concentra más bien en el puntaje verdadero estimado que en el puntaje obtenido. Su lectura es similar (para una mayor ampliación sobre éste tema véase el pie de pág. 205 del Manual de WISC). Digamos aquí que la asimetría se presenta en los puntajes extremos, pero no sucede lo mismo en la media (por eso en el ejemplo de Agus son bastante simétricos). Todo instrumento de medición es menos consistente en sus puntajes extremos que en los medios, por lo tanto el usuario debe tomar más precauciones en estos casos. A modo de ejemplo, veamos la siguiente tabla donde se presentan CI extremos y el CI medio.

Tabla 4.5. Intervalos de confianza de CI medio y CI extremos.

| CI | Intervalo 95% | Intervalo |
|---|---|---|
| 90% | | |
| 70 | 66 – 76<br>(-4 / +6) | 66 – 77<br>(-4 / +7) |
| 100 | 95 – 105<br>(-5 / +5) | 94 – 106<br>(-6 / +6) |
| 160 | 153 – 162<br>(-7 / + 2) | 152 – 163<br>(-8 / +3) |

Debajo de cada uno de los puntajes que señalan el intervalo de confianza ha sido calculado y puesto entre paréntesis la distancia de cada valor con respecto al puntaje obtenido. Como se puede observar, las diferencias sólo son simétricas, iguales, cuando el puntaje es el promedio (100); mientras que en el caso de los otros puntajes obtenidos, que son extremos, la asimetría va a favor de la cercanía con el puntaje medio.

## 4.8 Confiabilidad de las diferencias

En la práctica profesional, con frecuencia, es necesario considerar dos tipos de diferencias, las *interpersonales*, cuando se comparan las puntuaciones obtenidas por dos sujetos diferentes y las *intrapersonales*, cuando se comparan los puntajes obtenidos por un mismo sujeto en dos o más variables psicológicas.

Si –como se ha sostenido– el puntaje obtenido a partir de la evaluación de la técnica no es igual a la puntuación verdadera, ¿cómo saber si una persona obtuvo más puntaje que otra?; ¿cómo saber si una persona presenta puntuaciones diferentes al ser evaluada en distintos atributos?, cómo saber, en definitiva, si las diferencias entre los puntajes obtenidos son debidas al error de medición o reflejan diferencias reales en el atributo evaluado. Estas cuestiones implican la necesidad de analizar si las diferencias existentes entre distintas puntuaciones se deben al azar o a características del sujeto en la/s variable/s de interés. Estos interrogantes llevan concretamente a la siguiente pregunta: ¿qué distancia deben presentar entre sí los puntajes obtenidos para que efectivamente den cuenta de una diferencia real en el atributo evaluado?

En principio, conviene tener presente que el error típico de la diferencia entre dos puntuaciones es mayor que el error de medida de cualquiera de las dos, puesto que esta diferencia se halla afectada por los errores aleatorios presentes en ambas puntuaciones (Anastasi y Urbina 1998).

El error típico de la diferencia entre dos puntuaciones, puede hallarse partiendo de los errores típicos de medida de las dos puntuaciones a comparar, mediante la siguiente fórmula:

$$S_{dif} = \sqrt{S_{e\ med1}^{2} + S_{e\ med2}^{2}} \qquad [7]$$

Donde $S_{dif}$ es el error típico de la diferencia entre dos puntuaciones, $S_{e\ med1}^{2}$ es el error típico de medición al cuadrado del test 1 y $S_{e\ med2}^{2}$ es el error típico de medición al cuadrado del test 2.

La utilización del valor obtenido cuando se calcula el error típico de la diferencia, es similar a la del error típico de medición. Si es necesario un nivel de certeza del 95% sobre la diferencia entre ambas puntuaciones, estas deben entonces estar separadas entre sí por un valor igual a dos errores típicos de la diferencia. Una separación de sólo un error estándar de la diferencia entre ambas puntuaciones permitiría un nivel de confianza sólo del 68% de que estas son realmente diferentes, o sea, que reflejen diferencias reales en la variable evaluada.

En el siguiente ejemplo se presenta el tema, en un test que evalúa Personalidad y que incluye la medición del grado de Innovación que presentan los sujetos. El error de las diferencias es igual a 10 puntos. Se necesita comparar los puntajes de dos sujetos en el contexto de una evaluación en psicología laboral, a fin de saber si realmente las diferencias representan niveles de Innovación distintos.

Si las dos puntuaciones a comparar están separadas por 10 puntos, tendremos un 68% de seguridad de que esta diferencia entre los puntajes obtenidos refleja una diferencia entre puntuaciones verdaderas. Si las dos puntuaciones están separadas por 20 puntos, estaremos un 95% seguros de que la diferencia entre ellas representa diferencias de puntuación verdadera. Por último, si la distancia entre ambas puntuaciones es de 30 puntos o más, estaremos un 99% seguros de que la diferencia entre ellas representa diferencias en la puntuación verdadera.

Otros ejemplos. Tal como ocurre con las diferencias entre los puntajes de CI y los puntajes índice del WISC-III, la interpretación de la puntuación obtenida en un determinado subtest, como especialmente alto o especialmente bajo, debe estar precedida de una consideración de la significación estadística de la diferencia observada (además de una estimación de las frecuencias de base de la población).

La tabla B.3. del Manual (Wechsler, 1994, pág. 305) presenta la diferencia mínima requerida entre un puntaje de escala obtenido en un subtest y el promedio que alcanzó el niño o adolescente en un grupo de subtests (Verbales, de Ejecución o Escala Completa), para que tenga significación estadística. Se presentan diferencias para los dos niveles de significación que considera esta técnica.

El WISC-III también provee este tipo de información subtest por subtest. Para Analogías, por ejemplo, sólo una distancia igual o mayor a 3 puntos puede indicar una diferencia real entre la puntuación de Analogías y el puntaje medio -el puntaje esperado para la edad-, en el nivel de confianza del 95%. Se trata, en este caso, de una comparación interpersonal. Por lo tanto, si un niño obtuvo un puntaje igual a 8 en Analogías, y la media esperada para su edad es 10, su puntaje no puede asegurarse (con un nivel de confianza del 95%) que sea inferior al logrado en promedio por sus pares.

Con frecuencia, analizar la diferencia entre las puntuaciones obtenidas por un mismo sujeto, en determinado par de subtests, puede ser también de interés. En el Manual (Tabla B.4) se encuentran los datos necesarios, siendo el procedimiento a seguir similar al empleado en el ejemplo anterior.

*Capítulo 5*

# *Construcción y adaptación de técnicas psicométricas*

Alicia Cayssials
Marcelo Antonio Pérez

**Contenidos temáticos**

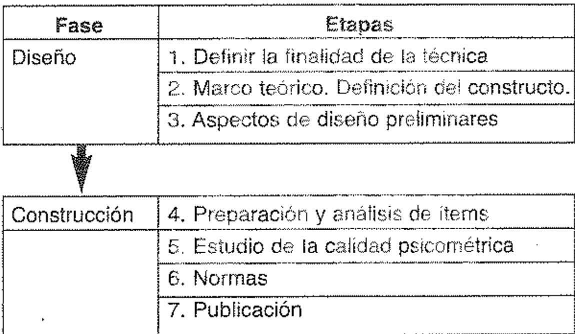
- **Pasos en la construcción de técnicas psicométricas**
- **Las diferencias culturales**
- **Adaptación de técnicas**

## 5.1 Pasos para la construcción de una técnica psicométrica

Este apartado pone el foco en los conceptos esenciales involucrados en el proceso de construcción de una técnica psicométrica. Dicho proceso es complejo, requiere de conocimientos exhaustivos tanto de la variable psicológica a evaluar como de las variadas técnicas de análisis de datos, y no es lineal –ya que el autor ronda de uno a otro buscando mejoras del instrumento en elaboración–. De todos modos, permite ser analizado, con fines didácticos, en dos fases: una primera de *diseño*, en donde se delinean los aspectos iniciales y basales del desarrollo del instrumento y en la que pueden distinguirse tres etapas, y una segunda de *construcción*, donde se materializa el instrumento, conformada por cuatro etapas. Estas dos fases y las etapas que las componen, conservan un orden lógico que es pasible de ser presentado sucesivamente, aunque en la práctica siempre habrá cierta yuxtaposición y realimentación entre etapas y fases.

El objetivo aquí es, entonces, trazar un recorrido que permita al lector la concatenación de las fases y sus correspondientes etapas y la integración de los contenidos desarrollados a lo largo de este curso básico.

El siguiente diagrama presenta en forma esquemática, las fases y etapas fundamentales que forman el proceso de construcción de un instrumento psicométrico, prescindiendo del tipo de teoría o de modelo psicológico que fundamente a la técnica.



| Fase | Etapas |
|---|---|
| Diseño | 1. Definir la finalidad de la técnica |
| | 2. Marco teórico. Definición del constructo. |
| | 3. Aspectos de diseño preliminares |
| Construcción | 4. Preparación y análisis de ítems |
| | 5. Estudio de la calidad psicométrica |
| | 6. Normas |
| | 7. Publicación |

El constructor comienza su trabajo de diseño en la primera etapa, mientras que el usuario comienza el suyo con el acceso a la publicación donde se presenta la técnica, es decir en la última etapa, cuando el instrumento ya está construido.

A través de esta publicación –que habitualmente es un Manual–, el autor de la técnica, en este caso quien ha elaborado el test psicométrico, brinda, además de la fundamentación teórica del instrumento, la información esencial necesaria para su aplicación, calificación y evaluación, el número y naturaleza de las personas en las que se establecieron las normas, así como los métodos utilizados para estudiar la confiabilidad y la validez. El usuario de una técnica psicométrica debe poder interpretar y valorar esta información.

No se encuentra dentro de los objetivos de este texto atender a los aspectos que deba realizar el constructor relacionados con la preparación de los materiales para su publicación (paso 7); tampoco se aborda aquí la etapa 6, ya que el capítulo 3 ha tratado el tema de la elaboración de normas y baremos. En los siguientes apartados se analizan, entonces, las cinco primeras etapas.

### Etapa 1. Definir la finalidad de la técnica

El análisis de la génesis de las distintas técnicas psicométricas permite establecer, en una primera aproximación general, dos tipos de propósitos en sus constructores. Por un lado se encuentran aquellos que elaboraron un test respondiendo a ***necesidades concretas*** de un ámbito de aplicación en particular. Un ejemplo clásico es la *Escala Métrica de Inteligencia*, elaborada por Binet y publicada por primera vez en 1905, cuya construcción fue realizada a partir de un problema y una solicitud del Ministerio de Educación de Francia (Zazzo et. al, 1970)

Por otro lado, se encuentran los investigadores que han elaborado instrumentos en el marco de ***desarrollos teóricos***. Ejemplo típico de este tipo de propósito es el que lleva a Lauretta Bender, en 1938, a elaborar el *Test Gestáltico Visomotor*, construido para realizar investigaciones sobre los principios postulados por la Teoría de la Gestalt. Bender crea el test con el objetivo de indagar la función gestáltica en niños y en pacientes con distintos trastornos psicopatológicos.

Más allá de cual de estos dos propósitos –u otros– movilicen al futuro elaborador de la prueba, este se encontrará inmediatamente ante una variedad de interrogantes



vinculados con la finalidad del instrumento, que deben responderse para guiar el proceso de diseño y construcción. Algunos de ellos, basados en los que describen Coehn & Swerdlik (2000), son los siguientes:

- ¿Cuál es el objetivo de la prueba?
- ¿Qué es lo que la prueba medirá de acuerdo a su diseño?
- ¿Cuáles son las necesidades de realizarla?
- ¿Hay otras pruebas que evalúen lo mismo? En caso afirmativo ¿Qué ventajas tendrá sobre otras y que desventajas?
- ¿Quién la usará? ¿Qué capacitación necesita para aplicarla?
- ¿A quién se aplicará? ¿Cuáles son las características de la población destino como ser su rango de edades, nivel cultural entre otras?
- ¿Qué beneficios les acarreará esta prueba?
- ¿Hay algún potencial daño que pueda ocurrir por la aplicación de esta prueba?

De modo más concreto, definir la finalidad de la técnica implica identificar las variables a medir y la población a la cual se dirige la evaluación. El primer interrogante del listado, ¿cuál es el objetivo de la prueba?, suele encontrarse expresado en los instrumentos que hay disponibles en el mercado, por medio de frases breves como las siguientes:

> El Inventario Multifásico de Personalidad Minnesota-2 (MMPI-2) es una prueba diseñada para evaluar la presencia de desajustes psicopatológicos en población adulta.
> El Inventario Clínico Multiaxial de Millon-III (MCMI-III) es una técnica construida para obtener información relevante sobre los trastornos de la personalidad.

El usuario de técnicas psicométricas no debe desconocer el hecho de que dos o más técnicas pueden compartir el mismo propósito, pero ser muy distintas unas de otras. Es decir, varias técnicas pueden evaluar variables tales como inteligencia, ansiedad, depresión o estrés, y sin embargo diferir ampliamente tanto en la forma de ponderarlas como en los aspectos que enfatizan al evaluarlas.

En los ejemplos mencionados, el MMPI-2 permite obtener información relevante sobre los trastornos de personalidad tal como lo hace el MCMI- III, –ya que estos son desajustes piscopatológicos–, y lo mismo vale a la inversa. Pero no se debe ignorar que cada una de las pruebas remite a un marco teórico y a una definición del constructo diferente.

El conocer el propósito explícito de una técnica psicométrica es solo una primera aproximación que puede resultar engañosa al usuario que suponga, a través de estas definiciones generales, que realmente el objetivo del instrumento coincide con el que él mismo se propuso evaluar. El profesional tiene que atender y profundizar el marco teórico y la definición de la variable que fundamenta su construcción, ya que solo con esta información puede valorar el tipo de decisiones que podrá tomar con las puntuaciones obtenidas, lo que supone también saber elegir entre varias opciones cuál es instrumento más adecuado a sus objetivos.

## Etapa 2. Marco teórico. Definición del constructo

La elaboración de un instrumento científico, implica la perspectiva de un marco conceptual que aporta información para la interpretación de las puntuaciones. Dicho marco no puede quedar reducido a un conjunto de definiciones de conceptos. El constructor de una técnica no solo tiene que definir los conceptos implicados, sino las relaciones lógicas de éstos con un marco teórico o una corriente psicológica más amplia.

El atributo psicológico en cuestión no puede captarse por sí mismo sin la mediación de un proceso intelectivo que dé cuenta de su sostén teórico. La base empírica que toda técnica psicométrica aporta, debe ser interpretada a través de la teoría.

Quien elabora una técnica ha de utilizar una serie de conceptos que tendrán un grado considerable de abstracción. Se trata de términos que deben ser definidos y relacionados entre sí a partir de la elección de una perspectiva teórica.

Desde el punto de vista del usuario, es necesario aclarar que no todas las publicaciones incluyen acabadamente estos aspectos. Hay técnicas que demandan encuadres teóricos de considerable desarrollo, mientras que otras solo enuncian el encuadre y proponen breves definiciones. Esta instancia depende, en gran medida, del grado de avance del conocimiento existente en el que se ha incluido el problema a indagar. De todos modos, es importante siempre que en la valoración de un manual de un test o de una publicación se tome en cuenta el relevamiento bibliográfico que ha realizado el autor.

En otras palabras, el constructor de un instrumento científico debe dar cuenta de la representación teórica del constructo que pretende evaluar. La definición de la variable a medir evitará la omisión de aspectos importantes del atributo o la inclusión de otros poco relevantes. Por ejemplo, en esta etapa debe quedar establecida la unidimensionalidad o la multidimensionalidad de la variable en cuestión, es decir, si se está midiendo un atributo que se considera pasible de ser representado unitariamente por un número, o por varios.

Por ejemplo, el Test de inteligencia WISC III propone evaluar el constructo inteligencia a través de un único número, el Coeficiente Intelectual, que está conformado por dos componentes, uno verbal y otro de ejecución, que dan lugar a sendos números, el Coeficiente Intelectual Verbal y el de Ejecución. Por su parte estos últimos también están compuestos por otras habilidades que pueden ser también medidas independientemente por otros números; la combinación adecuada de estos últimos dan como producto los coeficientes indicados. Es decir, el constructo inteligencia así operacionalizado para su medición, es un constructo multidimensional.

Desde el punto de vista del usuario, esta información le permite valorar el tipo de interpretaciones que podrá realizar a partir de la administración de la prueba.

## Etapa 3. Aspectos de diseño preliminares

Si el constructor de una técnica ya cuenta con un propósito claro, un marco o una perspectiva teórica que fundamenta la definición del constructo a evaluar, está en condiciones de comenzar la siguiente etapa que tiene como objetivo especificar "a priori" las principales restricciones con las que deberá operar el instrumento, tales como el tiempo de administración, los materiales a emplear, las situaciones y características de los sujetos a los cuales está destinado el test.



Los tipos de pruebas psicológicas son tantos y con propósitos tan diversos –verbales y motores, de rendimiento y de ejecución, de inteligencia, de aptitudes, de rendimiento, de personalidad, de intereses, de actitudes, que es imposible presentar un listado exhaustivo de los temas a considerar en esta etapa. A modo de ejemplo, a continuación se presentan los tópicos principales a ser considerados.

a) Tipo de test: basados en criterios o en normas.
b) Formato: Escala, cuestionario, inventario, entrevista.
c) Tipo de consigna: oral o escrita, explicaciones, necesidad de ejemplos o entrenamientos.
d) Tipo de respuesta: dicotómica, likert, diferencial semántico.
e) Características de los sujetos a examinar: edades, nivel de instrucción, nivel de comprensión lectora, nivel intelectual.
f) Modalidad de administración: indidivual, colectiva, autoadministrable, interactiva por computadora.
g) Tiempo de administración: con o sin tiempo límite, una única sesión o varias. En el último caso se tendrá que contemplar la forma de interrumpir y de continuar.
h) Forma de aplicación: oral o escrita, de lápiz y papel, manipulativo o de ejecución.
i) Tipo de exigencia: velocidad o potencia, grado o dificultad
j) Evaluación: manual o computarizada

Por otro lado, las decisiones relacionadas con el formato de los ítems, implican una tarea clave y compleja en el proceso de operacionalización del constructo, que comienza a realizarse en esta fase de diseño pero que se plasma en la práctica en la etapa siguiente.

## Etapa 4. Preparación y análisis de ítems

A partir de esta etapa comienza la fase de construcción del instrumento propiamente dicha, es decir que la definición teórica del constructo debe derivar en una operacional, entendiendo a esta como una definición concreta de la variable psicológica a evaluar, lo cual implica una revisión de las manifestaciones del constructo susceptibles de ser medidas. El constructor de una técnica tiene que exponer claramente las relaciones existentes entre la variable y sus manifestaciones observables, es decir, debe justificar que las respuestas solicitadas a los sujetos garantizan una medida relevante de la variable en cuestión (ver cap. 1).

Identificados los indicadores prácticos del constructo a medir a través de las definiciones operacionales, se hace necesario generar los estímulos que los fomenten, es decir los ítems cuya respuesta sea una manifestación observable de la variable a medir. Todas las técnicas psicométricas presentan *estímulos* o ítems a los que los *sujetos* dan una *respuesta; estímulos, respuestas y sujetos*, son importantes para conocer la naturaleza del instrumento (Cortada de Kohan, 1999).

Sea cual fuere la variable que mida, una técnica psicométrica está formada por un cierto número de elementos llamados reactivos o ítems, que aplicados al examinado fomentan en éste algún tipo de comportamiento como respuesta, vinculado al constructo que se desea medir. Su preparación y análisis constituye una etapa muy importante, debido a que la calidad de cada uno de ellos contribuye a la calidad del test en su totalidad.

Las tareas implicadas en esta etapa son variadas y complejas, y por este motivo pueden ser subdivididas en, por lo menos, cuatro momentos:

a) Confección de los ítems, consigna y formato de respuesta
b) Estudio pre-piloto
c) Administración y evaluación de la versión prelimitar en una muestra piloto de sujetos
d) Construcción de la forma definitiva del instrumento.

Entre estos pasos –al igual que entre las distintas etapas del proceso de elaboración de la técnica en general–, existe también un orden lógico, no necesariamente cronológico, por lo que la información obtenida en cualquiera de ellos puede ocasionar cambios en el plan original: en la práctica siempre habrá cierta yuxtaposición y un ir y venir sobre los distintos momentos, tanto de la etapa de diseño como de construcción.

El siguiente diagrama presenta los pasos en los que se sugiere dividir esta etapa y se señalan las interacciones más habituales entre ellos.

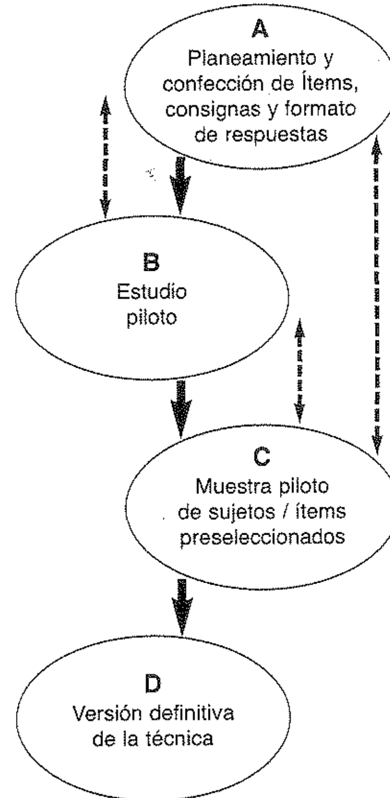



Fig. 5.1. Los cuatro pasos dentro de la etapa de preparación y análisis de los ítems. Las flechas negras indican el proceso lógico, la punteadas la posibilidad de avances y retrocesos dentro de los pasos.



En los tres últimos pasos de esta etapa, los ítems se ponen a prueba a través de administraciones de la técnica a los sujetos, es decir, los ítems se analizan a partir de la evidencia empírica sobre su funcionamiento. Por el contrario, el primero –planeamiento y confección–, es un paso que implica conocimientos de dos índoles: teórico-conceptuales y técnico-metodológicas.

*Paso A Planeamiento y confección de los ítems*

Muchos creen que la elaboración de una técnica comienza con la confección de los ítems, pero esto no es así. Sin una idea precisa sobre el propósito de la prueba, sin un adecuado ajuste a un marco teórico y una clara definición de la variable, no es posible abordar científicamente la construcción de un instrumento de evaluación psicológica. Es recomendable que la confección de ítems esté guiada por la teoría, es decir que evalúe variables que guarden un correlato con lo que esta plantea, ya que solo así se podrá elaborar un test con validez de constructo (ver Cap. 1).

En términos estrictos, la construcción propiamente dicha del test comienza al diseñar un conjunto numeroso de ítems, generalmente mucho mayor que la longitud prevista de la técnica (Martínez Arias, 1995). Este proceso implica determinar las posibles manifestaciones de la variable y debe ser realizado por expertos, ya que supone una discusión teórica; cabe aclarar al respecto, que muchas veces el experto que redacta los ítems no es el especialista en psicometría, siendo este último quién debe guiar al primero.

La confección de ítems es una parte muy importante del proceso. Es fundamental el respeto de los supuestos teóricos que enmarcan la definición del constructo que la futura técnica ha de evaluar, por lo que los ítems deben recoger información adecuada y representativa del atributo (Santisteban Requena, 1990).

En principio, el constructor de la técnica debe especificar un plan para el desarrollo de esta etapa en su conjunto y explicitar los criterios con los que llevará a cabo la selección y/o diseño de los ítems que serán incluidos en el instrumento. Luego debe identificar los observables representativos del constructo, del atributo de interés, y, si este tiene varios aspectos deberá establecer las proporciones de ítems que evaluarán cada uno de ellos: es importante que la magnitud aportada por cada uno de los ítems a la medición del constructo guarde la mayor correspondencia posible con su relevancia dentro del mismo. Así, los ítems que a partir de la teoría o de la experiencia práctica han sido hallados como más prototípicos de la variable a medir deberán tener una mayor representación en el puntaje final, que aquellos que son más accesorios. Cada ítem de una técnica psicométrica debe ser diseñado para inferir, a través de la repuesta que den los sujetos, una justa medida del atributo a evaluar, ya que este es el primer resguardo de su validez teórica y de contenido.

En los casos en que el constructo fuera multidimensional, debe informarse que dimensiones tiene la variable y que conjuntos de ítems componen cada una de ellas, además de especificar si la técnica arrojará un puntaje único y/o varios (por ejemplo, uno para cada dimensión). En el caso de que se ofreciera un puntaje único, producto de la combinación de varios, es necesario fundamentar como es que se integran los puntajes parciales para lograrlo.

Las decisiones sobre la naturaleza de los materiales estímulo y las posibilidades de respuesta del examinado, deben también fundamentarse y justificarse en virtud de los objetivos de evaluación que previamente se ha planteado. Un instrumento puede evaluar un determinado constructo –Ansiedad por ejemplo– pero deben precisarse a que

niveles de dicho constructo –en caso de tenerlos– se desea focalizar la evaluación –alto, medio o bajo por ejemplo–. Estas decisiones implican determinar el dominio que tendrá el instrumento, es decir precisar los niveles, rangos de medición y poblaciones objetivos de la evaluación y teniendo en cuenta estos alcances también se deberán diseñar los procedimientos de puntuación implicados.

Antes de que los especialistas en el tema comiencen a producir e inventar un gran número de ítems, debe seleccionarse el formato y el método de escalamiento que se usará en las respuestas; para los detalles a tener en cuenta en estos tópicos si sugiere ver el cap. 1 y 2 de esta obra. Para la elección del formato de las respuestas, conviene tener en cuenta la diferencia que hay en el diseño de ítems, en pruebas que miden:

a) inteligencia o aptitudes
b) aspectos de personalidad, intereses y actitudes

Las primeras tienen como objetivo evaluar el rendimiento de los sujetos, y para ello se utilizan técnicas de recolección de datos muy variadas. Los ítems pueden tener distintos grado de dificultad y la evaluación implica, en términos generales, valorar la adecuación de la respuesta del sujeto respecto a la establecida como adecuada, aunque también puede evaluarse el tiempo de ejecución y el tipo de error cometido entre otras posibilidades. Entre las múltiples decisiones que debe tomar el constructor de un test de este tipo, está la elección del tipo de respuesta que se le solicitará al sujeto examinado, si éste debe "construirla" o seleccionarla a partir de alternativas entre otras alternativas. Algunas opciones típicas de ítems de este tipo de instrumento pueden ser:

**Ítem con opciones de respuestas** (el sujeto debe seleccionar una respuesta)

¿Qué acontecimiento se conmemora el 17 de agosto?
☐ 1: Fallecimiento de San Martín
☐ 2: Día del trabajo
☐ 3: Batalla de Caseros
☐ 4: Primera invasión inglesa

¿Cuál es el resultado de multiplicas 40 X 60?
☐ a: 46
☐ b: 2400
☐ c: 2460
☐ d: 46000

**Ítem sin opciones de respuestas** (el sujeto debe construirla)
¿Qué acontecimiento se conmemora el 17 de agosto? ..............................................

*Nótese que en ambos casos hay una respuesta correcta (puede ser más de una) y otras que no lo son, pero cuando no hay opciones de respuesta, la posibilidad de adivinarla por parte del examinado es mucho menor.*

Por el contrario, el segundo tipo de técnicas, utiliza frecuentemente el formato de cuestionarios e inventarios que evalúan valoraciones de los sujetos, respecto a características o conductas personales, ideas o creencias con que se describen. Los ítems



que los conforman se caracterizan porque suelen realizarse comenzando por una breve descripción de alguna de dichas valoraciones y luego presentar dos o más respuestas alternativas, solicitando al sujeto que categorice o valore un orden.

A continuación se ejemplifican algunos de los formatos de respuesta más usuales en este tipo de instrumentos.

**Ítem dicotómico** (admite dos respuestas que, a su vez, son excluyentes)
A menudo estoy triste    **SÍ**    **NO**

**Ítem con tres alternativas** (el sujeto debe seleccionar solo una)
Si estoy muy enojado

| | | |
|---|---|---|
| **A.** Intento calmarme | **B.** Sigo así hasta que se me pasa | **C.** Me voy irritando cada vez más |

**Ítem con criterios de valoración** (el sujeto debe seleccionar solo una)
Estoy conforme con la carrera que elegí
☐ 1: no
☐ 2: algunas veces
☐ 3: muchas veces
☐ 4: casi siempre
☐ 5: siempre

Otro tipo de alternativas de valoración para último ejemplo podría ser:
Estoy conforme con la carrera que elegí
☐ 1: muy en desacuerdo
☐ 2: en desacuerdo
☐ 3: indeciso/a
☐ 4: de acuerdo
☐ 5: muy de acuerdo

Respecto al escalamiento de las respuestas, es de particular importancia la selección del nivel de medición, tanto de cada uno de los ítems como de la variable en su conjunto, ya que esto define las posibilidades que tendrá el tratamiento posterior de los resultados (ver cap. 1 y 3). También se ha visto en esos capítulos que la asignación de numerales a las respuestas da lugar a cuatro tipos de escalas (nominal, ordinal, de intervalos iguales y de cocientes), y que de acuerdo al tipo de escalamiento implementado es que se podrán o no realizar distintas operaciones matemáticas o lógicas entre los números asignados a las respuestas. Si bien en general es deseable que el nivel de medición sea el más elevado, se recuerda que solo se podrá usar aquel que respete el isomorfismo entre el numeral y la variable. (ver cap 3).

En general los instrumentos que miden variables de inteligencia y aptitudes tienen facilitado el proceso de cuantificación, en virtud de que al tener posibilidad de construirse con reactivos cuya respuesta puede determinarse si es correcta o incorrecta o en que gado lo es, el puntaje que se asigne a cada opción no tiene el nivel de subjetividad que caracteriza a los instrumentos dedicados a evaluar aspectos de personalidad, intereses y actitudes. La asignación de puntajes a las respuestas de los primeros, tradicionalmente se realiza otorgando números mayores a niveles de acierto mayores. En el ejemplo que pregunta cual es el resultado de la multiplicación de 40 X 60 hay un solo resultado correcto (2400), pero las opciones de respuesta ***a*** (100) y ***d*** (46000) tienen un

error evidentemente más grosero que la ***c*** (2460), mucho más próxima al resultado correcto, por lo cual podría darse distinta valoración a cada una de esas respuestas. Es decir, se podría dar un punto a la respuesta ***b*** y cero a las otras, o bien dos puntos a la ***b***, uno a la opción ***c*** y cero a las otras. Finalmente, cuando una escala está conformada por muchos ítems el puntaje total de la misma generalmente se obtiene con una simple sumatoria de esos números con que se valoró cada ítem que la compone.

En el caso de los instrumentos que evalúan aspectos de personalidad, intereses y actitudes, la puntuación suele hacerse de una forma similar. Por ejemplo en el caso del ítem dicotómico ejemplificado arriba (SÍ/NO), se puede asignar un punto a la respuesta si y cero a las respuesta no. Si varios de estos ítems conformaran, por ejemplo, una escala –supongamos de Depresión–, se podrían contar las frases en que el sujeto ha reconocido síntomas de dicho constructo y esa cantidad sería el puntaje obtenido.

Cuando hay criterios de valoración éstos, por sí mismos, suelen conformar un escalamiento ordinal, como el de los ejemplos dados más arriba, donde el responder con la opción ***5*** indica o mayor nivel de acuerdo (o mayor frecuencia), por lo que puede usarse este mismo número como indicador del nivel de acuerdo del examinado( o frecuencia). Si varios de estos ítems conforman una escala, se combinan los resultados de ellos para obtener un único número, frecuentemente a través de una simple suma. Para más detalles de cómo asignar números a los indicadores, ver el capítulo 3.

Teniendo ya definido el tipo de ítem y su formato de respuesta, el especialista procede a generar gran cantidad de ellos, que como se dijo anteriormente, suele ser mucho mayor al que finalmente conformará la prueba. La siguiente tarea implica someterlos a un minucioso análisis crítico o a un sistema de jueces o a ambos.

Si bien los reactivos fueron desarrollados por expertos y psicómetras, es menester que puedan superar el examen crítico y pormenorizado sobre su calidad y ajuste por parte de otros jueces, que puedan determinar su adecuación a la teoría e indicar cuales son los más adecuados y los que deberían desecharse. Si bien una vez finalizado el instrumento se harán los correspondientes estudios de validez y confiabilidad, la validez conceptual, de contenido y aparente en buena medida comienzan a ser evaluadas en este preciso momento. Sólo aquellos ítems que pasen por la aprobación de los jueces (generalmente se eligen en número impar para desempatar) integrarán el instrumento.

Además del conjunto de ítems, el formato y la valoración de las respuestas, también se deberá agregar la consigna general y/o las particulares si las hubieren, el procedimiento de corrección, los cuales también pasarán por la vista de los jueces, complementando la estandarización de una primera versión de la técnica que, como se verá en los pasos siguientes, será administrada a una *muestra piloto* o, lo que es más aconsejable, pasará a ser analizada en un *estudio pre-piloto*.

Antes de pasar al próximo apartado, es necesario aclarar que además de los ítems que evalúan el constructo en cuestión, pueden agregarse al instrumento otros con distintos fines. Como se dijo anteriormente, en los instrumentos que evalúan personalidad, actitudes e intereses, la subjetividad de la respuesta es mucho mayor que en los de aptitudes, y por ello también estas son susceptibles al falseamiento por parte del examinado, sobre todo en situaciones de evaluación obligatorias, como sucede en los exámenes laborales y forenses. Por ello es que aquí se hace especialmente importante controlar las disposiciones y el estilo de respuesta de los examinados, como ser la tendencia a describirse en forma distorsionada, teniendo en cuenta más lo que estos suponen que deben responder para lograr favorecerse del examen, que lo que piensan sobre como son o que les sucede en realidad.



Para evaluar estas tendencias se deben desarrollar indicadores que permitan determinar la distorsión en el estilo de respuesta, que suelen ser índices o escalas –denominados "de validez"– y que muy comúnmente en este tipo de instrumentos se obtienen a través de estos ítems, distintos de los diseñados para los fines de la evaluación de la variable psicológica en cuestión. Entre las distorsiones más frecuentes que se tratan de evaluar con estos índices están las tendencias a falsear las respuestas (mentir, simular, dar una imagen favorecía o desfavorecida de sí), la tendencia a la aquiescencia (por ejemplo a responder todo por "sí" en un inventario dicotómico) y las inconsistencias (responder de distinta manera ítems que evalúan lo mismo o casi lo mismo). Los índices o las escalas que se desarrollan con estos ítems suelen tener características formales similares a las de los constructos psicológicos, pudiéndose expresarse con puntajes brutos o transformados, o utilizarse criterios.

Como estos ítems no son específicos del constructo a medir, no siempre es necesario que los desarrollen expertos en ese tema, sino especialistas en psicometría, ya que en general están basados en ponderaciones estadísticas, como suele ser la infrecuencia de que determinado conjunto de respuestas sea respondido de una determinada manera. Por tal motivo es que no siempre se realizan en esta etapa sino que suelen trabajarse e intercalarse en la etapa final de construcción del instrumento.

En el caso particular de las escalas de inteligencia o aptitudes no es tan frecuente la inclusión de ítems de validez, ya que puede recurrirse a otros indicadores de consistencia mucho más sencillos, que evitan agregar reactivos que lleva tiempo desarrollar y además alargan la técnica. Así por ejemplo, si los ítems están ordenados por nivel de dificultad, el hecho que haya respondido bien a los finales y no lo haya hecho tanto a los iniciales indica una dificultad inusual que amerita justificarse, y que fácilmente puede ser expresada en un índice. Un ejemplo clásico de este tipo de índice es el llamado Discrepancia en el Test de Matrices Progresivas de Raven (Raven, 1993), en el cual se comparan los resultados obtenidos por el sujeto (aciertos) en cada una de las series que componen dicha técnica con el resultado "normal" (lo que el común de los sujetos con un desempeño similar al del examinado acierta en cada serie). En caso de que el examinado tenga un nivel de discrepancia muy elevado deberá revisarse el porqué de su desempeño tan irregular respecto al esperable.

*Paso B Estudio prepiloto*

Se trata de un ensayo, generalmente realizado en pequeños grupos de sujetos similares a los que está destinada la técnica, y que tiene como objetivo identificar ítems débiles o defectuosos, elementos con significado ambiguo, así como estimar la adecuación del lenguaje y las dificultades de comprensión.

En términos generales, el objetivo básico es analizar, con cierto detalle, el contenido de la técnica, y por ello con frecuencia se anexa un cuestionario donde se solicita a los sujetos que realicen comentarios sobre esta en general y sobre el comportamiento de los reactivos en particular. A su vez, los examinadores a cargo, registran puntualmente los efectos de las condiciones de administración, los problemas y las dificultades observadas.

En ocasiones es conveniente valorar el proceso implícito solicitando a los integrantes de los grupos que verbalicen cómo han llegado a la respuesta, es decir, se debe garantizar al máximo que el proceso de decisión del como responder esté relacionado con la variable y no contaminada por la influencia de otros factores improcedentes para la técnica en cuestión.

Un ejemplo hipotético pero plausible permite estimar la importancia de estos estudios. En un estudio prepiloto de una técnica destinada a niños, el constructor puede observar, con cierto grado de sorpresa, que la palabra "ratón", en la actualidad, en un grupo importante de chicos, se halla más asociada a un accesorio de la computadora que a un animal, lo cual aporta una significación totalmente distinta al ítem que había preseleccionado.

El estudio prepiloto permite desechar o corregir los ítems que habían sido incorporados a la primera versión, considerándose una buena práctica en el desarrollo de los instrumentos. Como resultado de este paso queda conformada la versión del instrumento, que se probará en "el campo" en el próximo paso: llamaremos a este instrumento versión piloto.

*Paso C Muestra piloto de sujetos / Ítems preseleccionados*

La elaboración de una técnica psicométrica es un proceso que lleva mucho tiempo, paciencia y dedicación. Los mejores resultados se obtienen con el trabajo en equipo y varios ensayos en muestras piloto, todos los que sean necesarios para depurar el instrumento.

El diseño de esta actividad implica una estimación previa del tamaño y de la delimitación de las precisas características de la población a la cual está dirigida la técnica, como los criterios de decisión sobre el número y composición de la muestra representativa de la misma. También requiere la versión piloto del instrumento, ya obtenida en el paso anterior, que irá modificándose sucesivamente a partir de los hallazgos que se logren en estos ensayos.

Básicamente se trata de administrar la versión piloto a la muestra representativa de la población a la que va dirigida el instrumento, de forma tal de evaluar el funcionamiento del mismo y obtener un conjunto de resultados concretos que permitan cuantificar y cualificar las características de los ítems. A partir de esos resultados, se pueden tomar decisiones ya fundamentadas en la práctica, que incluyen la modificación, inclusión o exclusión de ítems, modificación de la/ consigna/s, entre otras posibilidades.

El conjunto de procedimientos formales para hallar esta información se conoce en la literatura psicométrica bajo la denominación ***"análisis de ítems"***, siendo estos muy variados en función tanto de la técnica en cuestión como de las características de sus resultados, ya que de ellas dependen los procedimientos de análisis estadístico a utilizarse. Así, por ejemplo, para poder analizar el valor promedio de un resultado, es necesario que la escala de la que se lo obtiene sea intervalar o de cocientes. Es decir no tiene sentido hallar un valor promedio de una escala de otro tipo como una nominal u ordinal, ya que en el primero el número solo representa un nombre (ninguna cantidad) y en el segundo las cantidades son arbitrarias, es decir que el valor promedio también lo será, cualidad que heredarán todos los procedimientos de análisis de ítems que se fundamenten en él.

A través del análisis de ítems se pueden obtener numerosos índices que facilitan la visualización de las propiedades de los ítems, algunos de los cuales son generales para todas las técnicas y otros específicos. Por ejemplo, los índices que tratan de evaluar la dificultad de un ítem o su posibilidad de ser acertado por azar solo son aplicables a técnicas que evalúan inteligencia o aptitudes, donde, como se dijo anteriormente, existen respuestas correctas e incorrectas

Dentro de los índices generales más usuales para el análisis de los ítems se destacan los destinados a evaluar el poder discriminativo del ítem, y aquellos que describen el grado de relación entre la respuesta al elemento y algún criterio de interés, sea éste interno o externo al propio test (índices de discriminación, de homogeneidad, de confiabilidad y validez del ítem).



A continuación se describen algunas de las características de ellos:

***Poder discriminativo del ítem:*** implica estudiar

(a) si capta diferencias entre los sujetos y

(b) si la diferencia medida se debe a diferencias reales en el constructo a evaluar (o se debe a la influencia de variables improcedentes).

El propósito del test en su totalidad y de cada ítem en particular es proporcionar información sobre las diferencias individuales en el constructo que el test pretende medir, por lo que la utilidad del test se maximiza cuando más elevado es el poder de discriminación que tienen sus ítems.

Para hallar el poder discriminativo del ítems se suele recurrir a los *índices correlacionales de discriminación*, que constituyen un grupo de índices que se basan en la correlación entre la puntuación alcanzada por el sujeto en el ítem y la puntuación total en el criterio (ver cap. ). El constructor de una técnica debe seleccionar el coeficiente de correlación adecuado para el cálculo en función de la naturaleza de las puntuaciones del ítem y del criterio; una vez elegido, se procede a calcularlo para cada ítem, siendo que cuanto mayor sea dicho coeficiente mejor será la discriminación del item. Los elementos con bajo poder de discriminación, normalmente se eliminan.

***Sesgo de los ítems:*** se considera que un ítem está sesgado cuando arroja puntuaciones significativamente diferentes en grupos específicos de examinados que, teóricamente, forman parte de la misma población a la que se va a aplicar el test (Santisteban Requena, 1990). Como se verá en el apartado 5.3.3., este tema ha sido estudiado fundamentalmente en las investigaciones acerca de las diferencias relacionadas con la etnia, aunque también puede evaluarse en relación con otras diferencias entre grupos, tales como la clase social, género, edad, región, hábitat o cualquier otra característica socio-demográfica de los sujetos.

La existencia o no de sesgo se establece determinando si los parámetros de los ítems varían o no a través de los subgrupos.

***Dificultad del ítem:*** en el caso de pruebas donde interesa evaluar aptitudes, los ítems han de elegirse teniendo en cuenta su dificultad para ser respondidos adecuadamente. El grado de dificultad se puede calcular en forma sencilla teniendo en cuenta cuantos sujetos en la prueba piloto han dado respuesta acertada al ítem, aunque para su cálculo interesa tener también en cuenta cual era la probabilidad de haber sido acertado al azar. Nótese al respecto que en el ejemplo dado *¿Cuál es el resultado de multiplicas 40 X 60?* al ofrecer cuatro opciones de respuesta ya hay 1/4 (25%) de posibilidades de acertar la respuesta correcta por adivinación, en cambio si solo se pide el resultado sin dar opciones, la probabilidad de adivinarlo es sensiblemente menor.

***Confiabilidad y validez de los ítems:*** Análogamente a como se procede con las escalas que conforman el instrumento, es factible calcular la confiablidad y la validez de cada uno de los ítems. La forma de llevarlo a cabo es similar a la explicada en los capítulos 3 y 4 para el instrumento en su conjunto, aplicando coeficientes de correlación adecuados a las características de la prueba piloto y de los ítems en cuestión. Se

calculan los índices de confiabilidad y validez para cada ítem y en virtud de ello se seleccionan los elementos con mayor nivel de calidad, ya que serán los que maximizarán la validez y confiablidad del instrumento.

***Relación entre los ítems:*** El análisis factorial es un método frecuentemente utilizado en el análisis de los ítems (v. cap. 2), y se trata, en términos sencillos, de un método que permite determinar el nivel de relación que existe entre las respuestas a los ítems, y con esos resultados identificar conjuntos de reactivos que tienen algo en común (que tienen respuestas relacionadas, con niveles de correlaciones ente sí elevados) a los que se llamarán factores. Este tipo de análisis permite encontrar evidencia empírica de hipótesis teóricas del funcionamiento de los ítems y depurar aquellos que tienen un bajo "peso" en el factor correspondiente, es decir que tengan una baja correlación con los otros reactivos que previamente se consideraban estar evaluando lo mismo.

Además de los descritos, son numerosos los índices de calidad de cada ítem que pueden calcularse para guiar las sucesivas revisiones de la técnica, con miras a producir un test definitivo que tenga confiabilidad y validez máximas. Dependiendo de los propósitos para los que se construye la técnica, quien lo elabora debe prestar mayor consideración a unos u otros.

En síntesis, las puntuaciones obtenidas en las administraciones realizadas en una o más muestras piloto, permiten establecer definitivamente cuestiones específicas de la administración (consignas, materiales, tiempo, número de elementos) y a su vez determinar objetivamente las características de los ítems que pasarán a constituir la versión definitiva del instrumento.

*Paso D Versión definitiva de la técnica*

Una vez seleccionados los ítems que se consideran idóneos para la formación del test, se estudian las características de éste y se aplican técnicas para su estandarización definitiva, que incluirá el formato, las consignas, cuales reactivos la compondrán (también su orden, intercalamiento de ítems, de validez) las normas y los estudios de calidad psicométrica que se verán en el próximo apartado.

Con normas nos referimos a aquellos valores que deberán calcularse para que el usuario final pueda lograr una correcta valoración e interpretación de los resultados de la aplicación (véase puntajes transformados en cap. 3). Para obtenerlos, el instrumento es aplicado a una muestra representativa de aquella población a la que va dirigido, la que se denomina *grupo normativo*.

En resumen, el tratamiento de los ítems es una de las etapas más complejas y largas, y que amerita el mayor cuidado por parte del constuctor. Cabe solo remarcar que, como se expresa en la figura 5.1, antes de realizase la versión definitiva se deben hacer todos los cambios y pruebas piloto necesarias hasta maximizar las cualidades evaluativas del test. Luego de ello se pasará a la etapa siguiente, en la que se evalúa la calidad psicométrica, cuyos estudios pueden determinar también la necesidad de volver a hacer algún ajuste de los correspondientes a la presente etapa.



Etapa 5. Estudio de la calidad psicométrica

Las dos cualidades de un instrumento psicométrico en las que el investigador y el usuario deben interesarse especialmente son la *confiabilidad* y la *validez*. El lector puede encontrar un análisis pormenorizado de ambas en los capítulos respectivos (véase capítulos 2 y 4).

Respecto a la confiabilidad, las pruebas que miden aptitudes suelen alcanzar una mayor que las que miden aspectos de la personalidad. Esto suele deberse a que, –como se dijera en el apartado anterior paso, A–, las respuestas de este tipo de pruebas pueden identificarse como correctas e incorrectas (son menos subjetivas) lo que hace, desde el punto de vista métrico, más sencillo evaluar una aptitud que una actitud.

En cuanto a los estudios de validez, generalmente en las pruebas de aptitud también es más sencillo encontrar criterios externos cuantificables, tales como el rendimiento académico y/o laboral. De todos modos, los constructores de estas técnicas suelen presentar distintos tipos de estudios para dar cuenta de la validez de las puntuaciones. Por su lado, las técnicas que evalúan características de personalidad, intereses o actitudes, en general, ponen el énfasis en la validez de contenido, de constructo, y con frecuencia analizan los puntajes a través del análisis factorial. Un diseño para estimar la validez también muy utilizado es el procedimiento de *grupos contrastados* (ver cap. 2).

No se dedica aquí más espacio a la importante etapa del estudio de la calidad del test, ya que fueron profundizados en sendos capítulos.

## 5.2 La adaptación de los test

El presente apartado delimita cuestiones relacionadas a la valoración de la adaptación de un instrumento, es decir a los estudios que deben realizarse para ajustar una prueba original proveniente de un determinado medio sociocultural y adaptarlo a otro. Si bien no desarrolla exhaustivamente las distintas tareas que se deben llevar a cabo para adaptar una prueba, proporciona un panorama general sobre el tema.

La elección de una técnica psicométrica requiere cautela y reflexión en cuanto a su adecuación cultural y a su actualización. Cuando un investigador *adecua* una técnica en uso desde el punto de vista de su ajuste cultural, realiza una ***adaptación*** del test en sentido estricto, mientras que cuando la *actualiza*, realiza también una adaptación, pero en este caso se denomina ***revisión***.

La adaptación cultural y la revisión están fuertemente imbricadas. Toda adaptación involucra siempre atender las especificidades de una comunidad así como al carácter cambiante de ésta. La cultura no es estable y permanente. En nuestros días, el proceso de interpenetración cultural, por ejemplo, está presente en todas las sociedades. Se trata de un factor dinámico a tener en cuenta tanto por los constructores como por los adaptadores y los usuarios de los tests.

Berry (2004) define cultura como *"un modo de vida compartido por un grupo de personas. Estos patrones de conducta (que incluyen cogniciones y emociones) son explícitos e implícitos. Están constituidos por símbolos, ideas y valores que se transmiten de una generación a otra"*.

Es importante remarcar que las diferencias culturales no se refieren a las que hay entre naciones ni etnias, ya que dentro de un mismo país, incluso dentro de una

misma ciudad, barrio o institución, pueden hallarse patrones conductuales que definen distintos grupos culturales y subculturas.

Durante las primeras décadas del siglo pasado, ya se hizo notorio que toda técnica de evaluación es el resultado de una cultura y que responde a los valores de la misma. El examen de inmigrantes, la evaluación psicotécnica de extranjeros para acceder a puestos de trabajo y distintas investigaciones señalaban que los tests tradicionales de inteligencia y de personalidad presentaban un sesgo cultural que los hacía inadecuados para la evaluación de minorías étnicas o sujetos de culturas diferentes a la dominante.

Las experiencias en distintos ámbitos de aplicación pusieron en evidencia las limitaciones de emprendimientos para crear técnicas denominadas *Libres de Influencia Cultural* (L.I.C.). Entre los instrumentos pioneros, es clásico mencionar el test de los Cubos, de Kohs, creado en 1914 y el test del Laberinto de Porteus, elaborado en 1924. Estos instrumentos, salvo en su consigna que era por lenguaje oral o escrito, requerían de una tarea casi exclusivamente manipulativa, y la consigna podía transmitirse en forma escrita, oral o por imitación, lo cual auspiciaba la posibilidad de que fueran panculturales, que pudieran resolverse en forma independiente de la cultura a la que pertenecía el examinado. Podemos sintetizar las críticas que surgieron ante estos intentos mencionando algunas afirmaciones de Anastasi (1998) respecto del tema:

- Resulta inútil intentar elaborar un test L.I.C..
- Ninguna prueba puede ser aplicable universalmente.
- Resulta improbable que algún test pueda ser igualmente imparcial con respecto a más de un grupo cultural, especialmente si las culturas son muy distintas.
- Cada test tiende a favorecer a las personas de la cultura en la que se ha creado.

En cuanto al intento de desarrollar Test no verbales, porque estos serían más universales, Anastasi (1998) agrega:

- No se puede dar por sentado que los tests no verbales midan las mismas funciones que los verbales.
- Las pruebas no verbales pueden estar más saturadas culturalmente que las verbales.
- Si las variables psicológicas resultan de la combinación de comportamientos importantes dentro de una cultura, ¿para qué eliminar las diferencias culturales?

Todas las técnicas tienen determinados prerrequisitos para que una persona pueda realizarla. El lenguaje utilizado, los modos de comunicación, el nivel de vocabulario necesario para la comprensión de la consigna, si los ítems le son familiares e interesantes, entre otros, son todas variables que tendrán algún correlato con el desempeño de los sujetos en la prueba, y por ende en los resultados obtenidos a través de ella. Si no se puede lograr un test libre de influencia cultural, lo deseable es que sea culturalmente justo, es decir que haya igualdad de posibilidades entre todos los grupos con él evaluados, sobre todo los minoritarios.

Mientras que valorar el grado de actualización de una técnica es relativamente sencillo, (para ello basta con detenerse en el año de publicación del instrumento y revisar su bibliografía); por el contrario, relevar el ajuste cultural de un instrumento es un proceso más complejo. Los siguientes apartados introducen conceptos claves para hacerlo.



*Emico y ético. El análisis de las equivalencias.*

El interés por el efecto de las diferencias culturales en la evaluación psicológica en general, y en la utilización de las técnicas psicométricas en particular, data de mucho tiempo atrás y en la actualidad están claras las cuestiones que debe atender quien realiza la adaptación de un instrumento.

Marín (1986), en su artículo *Consideraciones metodológicas básicas para conducir investigaciones psicológicas en América Latina*, invita a reflexionar sobre el hecho de tener en cuenta la necesidad de evitar la falsa suposición de que los métodos y las ideas desarrolladas en una cultura son igualmente válidos en otra. Estas consideraciones implican, en principio, la comprensión apropiada de las diferencias entre los aspectos Emic ( *émico)* y Etic *(ético)*, introducidas por Kennet Pike, para diferenciar la fonética (fonemics) de la fonología (fonetics) en los estudios del lenguaje (Pike, 1954). Se denomina *éticos* a aquellos constructos o aspectos de los mismos, ideas e instrumentos, que tienen y han demostrado características universales, mientras que los aspectos denominados *émicos* son aquellos vinculados o utilizables en sólo uno o en pocos grupos culturales

La consideración de este tema no implica juicios valorativos, su objetivo es destacar la importancia de analizar el grado de universalidad del constructo o instrumento en cuestión y de demostrarlo con datos empíricos. Así por ejemplo, un constructo como inteligencia tiene un importante valor ético, debido a su universalidad: en casi todas las culturas existe alguna concepción de la capacidad de los sujetos, pero también posee alguna valoración distintiva de cada cultura en particular (valor émico). Lo que normalmente medimos como inteligencia a través de un test de coeficiente intelectual puede ser muy útil en contextos académicos o de rendimiento que se corresponden a la cultura de las ciudades de los países de hemisferio occidental, pero posiblemente no lo sea –y por ende sea poco valorado– en otra cultura o subcultura donde las habilidades medidas por dicho test no sean relevantes para la adaptación o la vida cotidiana de los sujetos.

La ciencia tradicionalmente ha dado un valor destacado a la universalidad de los hallazgos, y por ello la importancia que le da a la valoración ética; no obstante, en los últimos años, numerosas investigaciones transculturales han ido demostrando la falacia del valor ético de muchas teorías y constructos que se suponían tales, rescatándoseles su valor émico. Es decir, que cuanto mayor valoración ética tiene un constructo más universal es, pero es muy poco probable que no tenga un componente émico, que al trasladarlo a otra cultura no haya que hacer conversiones y equivalencias para valorarlo o utilizarlo en su justa medida. En este sentido, quien adapta una técnica debe ser ante todo un "mediador" entre culturas (Frank, 1999), y debe considerar detalladamente las características del instrumento original para adecuarlas a la nueva cultura en cuestión.

Marín (1986), en el artículo mencionado, propone tres tipos de equivalencias a tomar en cuenta en esta adaptación de un constructo: las equivalencias conceptuales o de constructo, lingüísticas, y las métricas, a las que agregaremos una cuarta, la equivalencia de formato.

***Conceptuales:*** se refiera a si el constructo existe en la cultura donde se desea utilizar la técnica en cuestión, y en tal caso, si la forma de valorarlo es la misma que en la cultura de origen. Esta equivalencia nos lleva a preguntarnos por la validez cultural del constructo y del instrumento que lo mide, si, por ejemplo, el comportamiento

valorado como inteligente tiene los mismos indicadores –actitudes, conductas– en la cultura origen que en la que se adapta el test. Está claro que si las culturas a nivel verbal y no verbal son muy difíciles de comparar, las pruebas son muy difíciles de trasladar de un contexto cultural a otro (Berry, 2004). Es muy posible que para lograr una acabada equivalencia conceptual se deba recurrir a equipos de especialistas como lingüistas, antropólogos, sociólogos entre otros.

***Lingüísticas:*** se refiere a la redacción de los ítems y consignas, a su traducción y al empleo de términos que tengan significados iguales o lo más parecidos posible a los originarios. La equivalencia lingüística está muy hermanada con la conceptual, ya que la traducción del instrumento no es una actividad lineal, sino que deberá tener en cuenta los giros idiomáticos, la idiosincrasia, las creencias y los valores puestos en juego en los reactivos, entre otros aspectos. Para realizarla, se suele recurrir a las traducciones por consenso (varios expertos), a personas bilingües y luego se realizan las pruebas piloto necesarias para garantizar la correcta equivalencia. Las malas adaptaciones lingüísticas redundan en dificultades de comprensión de las consignas y con ellas que se produzcan distorsiones en las puntuaciones (por ejemplo más bajas) o la abundancia de falsos negativos, es decir reactivos contestados mal o directamente no contestados.

Como se indicó anteriormente, no hay que suponer que ambos tipos de equivalencias (las conceptuales y lingüísticas) solo se realizan cuando se pretende adaptar un instrumento de una lengua a otra. Por ejemplo en el inventario de Estilos de Personalidad de Millon MIPS (Millon, 1997) podemos encontrar el siguiente reactivo al que se debe contestar por verdadero o falso: *Nunca he dejado estacionado el auto por más tiempo del que un parquímetro establecía como límite.* Podrá imaginar el lector que la traslación del ítem desde el original norteamericano al ciudadano de Buenos Aires es mucho más adecuada y clara que la de este último lugar a los centenares de ciudades y pueblos de nuestro interior, en los que la palabra parquímetro es desconocida. Da por sentado también que el lector usa habitualmente un auto.

***Métricas:*** esta adaptación se refiere al calibrado, tanto al valor con que se pondera cada ítem (si se debe mantener, cambiar), como a la adecuación de las normas, la revaluación de los estudios de confiabilidad y validez y la revisión de la cantidad de factores que componen el instrumento entre otras posibilidades. Muchas veces se confunde a la equivalencia métrica con la realización de un nuevo baremo regional, actividad que forma parte de la misma, pero que además abarca un análisis métrico integral, que va desde el estudio del comportamiento del ítem hasta el del resultado.

Así por ejemplo, los instrumentos que tienen ordenados los ítems de acuerdo a algún criterio como puede ser la dificultad del mismo, deberán calibrarse para la nueva cultura, lo que implicará con toda seguridad cambios en el ordenamiento originario debidos a cambios en la dificultad de los reactivos. Téngase en cuenta la importancia de esto en aquellas técnicas, como por ejemplo en los Subtest de Vocabulario o Información del WISC IIII, donde hay puntos de comienzo y terminación que fueron calculados de acuerdo a la dificultad de los ítems que lo componen. No adaptar este aspecto puede implicar sobre-clasificar o sub-clasificar a los evaluados.

***Formato:*** Se refiere a los aspectos formales del instrumento que puedan afectar la forma de responder de los sujetos, como ser la utilización del tiempo, el tipo de formato de las respuestas.



Por ejemplo, el instrumento EDI-2 Eating Disorder Inventory- 2 (Garner 1991) es un inventario de 91 frases las que se pueden responder con seis opciones de respuesta muy frecuentes de hallar en instrumentos de habla inglesa: ***never, rarely, sometimes, often, usually, always.*** Cuando se adaptó el instrumento a la Argentina (Casullo, Pérez 2005), y en vista que en castellano no hay seis palabras equivalentes, se propusieron como tales las siguientes opciones:

| | | |
|---|---|---|
| ***Never:*** *nunca* | ***rarely:*** *pocas veces* | ***sometimes:*** *algunas veces* |
| ***Often:*** *bastantes veces* | ***usually:*** *muchas veces* | ***always:*** *siempre* |

Cuando se hizo la prueba pre-piloto se detectó que quienes la integraron tuvieron dificultades para distinguir las diferencias entre las opciones ***pocas veces*** y ***algunas veces***, o entre ***bastantes veces*** y ***muchas veces***. El anglosajón tiene muy internalizada la significación y la frecuencia que representa cada una de las seis palabras indicadas, por lo que le es familiar y natural responder con ellas, pero en nuestro medio esto no es así. También se presentaron dificultades para que los examinados optaran entre siempre o nunca, ya que los consideraban valores muy extremos, muy absolutos, por lo que se decidió reemplazarlos por casi nunca o casi siempre.

Tras sucesivas pruebas pilotos se fueron cambiando los formatos de respuesta hasta que se concluyó que era mucho más adecuado a nuestro medio la utilización de solo cuatro opciones: ***Nunca o casi nunca, a veces, muchas veces y siempre o casi siempre,*** formato con que quedó finalizada la adaptación de las opciones de respuesta.

Un aspecto a destacar cuando se hacen las equivalencias es el referido a cómo afecta el tipo de la respuesta llamada ***Deseabilidad Social***, entendiendo como tal la tenencia de los sujetos a contestar los ítems de acuerdo a lo que es "esperable" que responda desde una perspectiva de valoración social, y no en virtud de cómo realmente es o lo que le sucede. Es bien sabido que el uso de inventarios es muy frecuente en los países de habla anglosajona para realizar todo tipo de evaluación, como así también la confianza que allí se tiene sobre la utilización y confidencialidad de los resultados, por lo cual hay una predisposición y actitud bien distinta hacia la modalidad evaluativa que la que se puede encontrar en medios sociales como el nuestro. En general esta actitud suele producir un error sistemático (sesgo) en los resultados ya que la mayoría de los sujetos responde en un sentido del ítem –el favorable a deseabilidad social–.

Muchas veces, el sesgo originado en el efecto de la deseabilidad social se puede corregir realizando baremos del instrumento para la población específica, como pueden ser, por ejemplo, los baremos de instrumentos de evaluación de personalidad para ámbitos laborales o no laborales. Así, los postulantes a puestos laborales, durante los exámenes de admisión y en su afán de alcanzar el puesto, con facilidad tienden a responder los instrumentos dando una descripción favorecida de sí mismos. Por ello, en aquellas situaciones donde se considere que la deseabilidad social es una variable de peso, se debe examinar con mayor minuciosidad la adecuación de la muestra con la que se realizaron los baremos con que comparamos al examinado. Se han hallado en nuestro medio grandes diferencias entre las evaluaciones obligatorias –como las laborales– y las no obligatorias –con población general voluntaria– debidas al efecto de la deseabilidad social (Castro Solano, Casullo, Pérez, 2004), que han justificado la creación de normas diferenciadas para ambos grupos.

En síntesis, se requiere establecer la validez cultural del constructo, la validez lingüística y/o gráfica de su expresión en el test, la validez del formato seleccionado y la validez métrica. Esta no es una tarea simple, muchos de los tests, aún los más utilizados, evidencian cierto *sesgo*.

Para tratar de disminuir este sesgo, hacen falta múltiples investigaciones multimodales y/o paralelas que contrasten los resultados obtenidos en diferentes grupos poblacionales, y que tengan en cuenta los aspectos antes desarrollados.

### *Sesgo y equidad*

Cuando Binet encuentra, en 1910, que los niños de estatus socieconómicos más bajos rendían peor en algunos ítems de su test, pensó que los elementos de su escala, en esos casos, podían estar midiendo más los efectos del entrenamiento cultural que la capacidad mental.

Sin embargo, el problema del sesgo en los instrumentos de medida se ha convertido en un tema importante en la literatura psicométrica a partir de finales de los años setenta y surge ligado a la aparición de diversos movimientos por los derechos civiles en EEUU, con las reivindicaciones por la igualdad de derechos para algunos grupos considerados injustamente tratados en situaciones de selección para puestos de trabajo y de admisión en instituciones educativas, contextos en los que las decisiones se basan con mucha frecuencia en tests psicométricos (Martínez Arias, 1995).

En términos generales, el problema del sesgo apunta a la cuestión de si las diferencias entre grupos encontradas en los resultados de los tests reflejan diferencias reales en la variable medida entre los grupos o si éstas son causadas por fuentes sistemáticas de variación ajenas al constructo que mide el test. En otras palabras, el propósito de las investigaciones sobre el sesgo es separar las diferencias reales de las artefactuales (generadas por el propio instrumento de medida). Se debe estudiar, entonces, el funcionamiento diferencial del ítem en sujetos pertenecientes, por ejemplo a minorías lingüísticas o con condiciones discapacitantes.

## Resumen

En síntesis, cuando un usuario tiene que valorar una técnica, de cara a su posible utilización, debe plantearse, además de otras cuestiones específicas, los interrogantes listados a continuación.

1) ¿qué teoría psicológica fundamenta al instrumento?
2) ¿cómo se ha estudiado la validez de constructo?
3) ¿cuán actualizada es la bibliografía consultada?
4) ¿presenta consignas adecuadas para los sujetos a evaluar?
5) ¿cómo fueron seleccionados los ítems?
6) ¿se ha analizado el sesgo para ciertos grupos o poblaciones?
7) ¿cuáles son sus propiedades psicométricas? (confiabilidad y validez)
8) ¿cómo se puntúa y cuáles son las características del grupo normativo sobre el cual se estandarizó?
9) ¿quién/es, cómo y cuándo fue adaptado a nuestro medio?: ¿han sido realizadas correctamente las equivalencias conceptuales, lingüísticas, métricas y de formato?



Cualquiera sea la técnica de evaluación que se utilice, si es administrada a un sujeto culturalmente disímil a la cultura donde proviene la técnica, se cometerán errores y lo peor es que, frecuentemente, estos pueden pasar inadvertidos. Sólo la formación de quien esté a cargo de la interpretación de los datos aportados por el instrumento puede subsanar, en parte, esta cuestión.

En este punto, son dignos de destacar los lineamientos que la Asociación Americana de Psicología agrega, en 1989, al código que reglamenta la formación y actuación del profesional (Frank, 1999), que se listan abajo.

a) un amplio conocimiento cultural provisto por efecto de la práctica clínica y de la investigación comparada;
b) el reconocimiento de la propia raigambre cultural;
c) el respeto por los valores, las creencias y la visión del mundo de otras culturas;
d) la utilización del lenguaje preferido por el entrevistado y, en su defecto, el uso de un traductor entrenado;
e) el reconocimiento del impacto del contexto socio-económico-cultural y político sobre los problemas que trae el consultante y las intervenciones propuestas;
f) una activa actitud de defensa de los derechos del consultante en relación con situaciones de discriminación y racismo que afecten negativamente a su salud física y/o mental.

Numerosos investigadores transculturales invitan a viajar a países exóticos para comprender sus indagaciones y sus trabajos resultan enriquecedores. Más allá de la posibilidad de realizar estos viajes, es importante señalar que en la actualidad ha crecido la conciencia de la existencia de valores culturales heterogéneos no solo en distintos continentes y entre países que conforman un mismo continente, sino también entre regiones que integran una misma nación como entidad jurídico-política.

Por último, con respecto a una perspectiva de esta problemática en el contexto de nuestra realidad sudamericana, se recomienda la lectura de la conferencia *La evaluación psicológica: Modelos, técnicas y contexto sociocultural*, impartida por Casullo (1999) en Salamanca, en ocasión de la VI Conferencia Internacional de Evaluación Psicológica.

En resumen, tanto los constructores de técnicas psicométricas, como los adaptadores, revisores y usuarios, cuanto más profundicen los determinantes culturales de las variables psicológicas, mejor podrán integrar e interpretar la información aportada por los sujetos evaluados y reconocer, a su vez, las limitaciones de su trabajo.

# *Bibliografía*


ADEIP (1999). *Código de ética del psicodiagnosticador.* Rosario: Asociación de Estudio e Investigación en Psicodiagnóstico.

ADEIP (2000). *Pautas internacionales para el uso de los Tests. Adaptación argentina de las normas de la International Test Commission.* Buenos Aires: Asociación de Estudio e Investigación en Psicodiagnóstico.

Aiken, L. R. (1999). *Personality assessment methods and practices.* 3ª ed. Seattle: Hogrefe & Huber.

Albajari, V. L. (1996). *La entrevista en el proceso psicodiagnóstico.* Buenos Aires: Psicoteca.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement (NCME), (1974). *Standards for Educational and Psychological tests and manuals.* Washington D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2004). *Standards for educational and psychological testing.* Washington D.C.: American Educational Research Association.

American Psychological Association (1954). *Technical recommendations for psychological testsand diagnostic techniques.* Washington DC: American Psychological Association

APA/ American Psychiatric Association (1995). *DSM-IV. Manual Diagnóstico y Estadístico de Trastornos Mentales.* Criterios diagnósticos. Barcelona: Masson.

Amón J. (1980). *Estadística para psicólogos.* Madrid: Pirámide

Anastasi, A. & Urbina, S. (1998). *Tests psicológicos.* México: Prentice Hall.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Measurement in Education, 1*: 215-222.

Anguera, M. T. (1995). *Avances metodológicos en evaluación.* Comunicaciones. Madrid: Colegio Oficial de Psicólogos.

Aspinwell, L. G. & Staudinger, U. M. (Eds.). (2003). *A psychology of human strengths: Fundamental questions and future directions for a positive psychology.* Washington DC: American Psychological Association.

Avila Espada, A. (1989). *Evaluación psicológica clínica.* Madrid: Provisional.

Avila Espada, A. (1997). *Evaluación en Psicología Clínica. Vol. II: Estrategias cualitativas.* Salamanca: Amarú.

Bellak, L. (1992). Projective techniques in the computer age. *Journal of Personality Assessment, 58,* 445-463.

Bericat, E. (1998). *La integración de los métodos cuantitativo y cualitativo en la investigación social.* Barcelona: Ariel.

Bibliograf (1997). *Spes. Diccionario Abreviado latino- español, español- latino.* Barcelona: Bibliograf.

Bingham, W. V. (1937). *Aptitudes and aptitude testing.* New York: Harper.

Botella, J.; León, O. G. & San Martín, R. (1997). *Análisis de datos en Psicología I.* Madrid: Pirámide.

Buck, J. N. (1948). The H-T-P technique, a qualitative and quantitative method. *Journal of Clinical Psychology, 4,* 317-396.

Buck, J. N. (1992). *House-Tree-Person projective drawing technique (H-T-P): Manual and interpretative guide* (Revised by W. L. Warren). Los Angeles, California: Western Psychological Services.

Burns, R.C. (1982). *Self-growth in families: Kinetic Familiy Drawings (K-F-D) research and applications.* New York: Brunner Mazel.
Burns, R. C. & Kaufman, S. H. (1970). *Kinetic Familiy Drawings (K-F-D): An introduction to understanding children through kinetic drawings.* New York: Brunner Mazel
Campbell, D. T. & Fiske, D. W. (1959). Convergent and dicriminat validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*: 81-105.
Castro Solano, A.; Casullo, M.M. & Pérez, M. (2004). *Aplicaciones del MIPS en los ámbitos laboral, educativo y médico.* Buenos Aires: Paidós.
Casullo, M. M. (1996). *Evaluación psicológica y psicodiagnóstico.* Buenos Aires: Catálogos de la Secretaría de Cultura, Facultad de Psicología UBA.
Casullo, M.M. (ed.) (1999). *El inventario MMPI-2 en los ámbitos clínico, forense y laboral.* Buenos Aires: Paidós.
Casullo, M.M. (1999). La evaluación psicológica: modelos, técnicas y contexto sociocultural. *Revista Iberoamericana de Diagnóstico y Evaluscíón psicológica,* 1: [98-111].
Casullo. M.M. (2003). *Adaptación del Inventario MMPI-A.* Departamento de Publicaciones. Facultad de Psicología. UBA.
Casullo, M. M. (Ed.). (2003). *El bienestar psicológico en Iberoamérica.* Buenos Aires: Paidós.
Casullo M. M. (2007). *El inventario de síntomas SCL-90 R de L. Derogatis.* Buenos Aires: Departamento Publicaciones Facultad de Psicología UBA.
Casullo M.M. & Pérez M. (2003). *El Inventario de Personalidad "Big Five" (cinco factores).* Buenos Aires: Departamento Publicaciones Facultad de Psicología UBA.
Casullo, M. M.; Figueroa, N. B. L. de & Aszkenazi, M. (1991). *Teoría y Técnicas de Evaluación Psicológica.* Buenos Aires: Psicoteca.
Cattell, R.B. (1975) *16PF, Cuestionario factorial de personalidad.* Madrid: TEA
Cayssials, A. ( 1998). *La escala de inteligencia WISC-III en la evaluación psicológica infanto-juvenil.* Buenos Aires: Paidós.
Cliff, N. (1973). Psychometry. En B. B. Colman (Ed.). *Handbook of General Psychology.* New Jersey: Prentice Hall.
Cohen, R. J. & Swerdlik, M. E. (2001).*Pruebas y evaluación psicológica. Introducción a las pruebas y a la medición.* México: Mc Graw Hill
Coolican, H. (1997). *Métodos de investigación y estadística en psicología.* México: El Manual Moderno
Cortada de Kohan, N. (1994). *Diseño estadístico.* Buenos Aires: Eudeba
Cortada de Kohan, N. (1999): *Teorías Psicométricas y Construcción de Tests.* Buenos Aires: Lugar Editorial.
Cortada de Kohan, N. (2004.) *Teoría y Métodos para la Construcción de Escalas de Actitudes.* Bs. As.: Lugar.
Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.
Cronbach, L. J. & Gleser, G. C. (1957). *Psychological Tests and Personnel Decisions.* Chicago: Illinois University Press.
Cureton, E. E. (1950). Validity, reliability and baloney. *Educational and Psy chological Measurement,* 10:94-96.
Costa P. & McCrae R. (1992). *NEO-PI-R.* Odessa: Psychological Assessment Resources.
Derogatis, L. (1994). SCL-90-R. *Symptom Checklist-90-R. Administration, Scoring and Procedures Manual.* Minneapolis: National Computer Systems.
Exner, J. E. Jr. (1995). *Issues and methods in Rorschach research.* Mahwah, New Jersey: Erlbaum.
Fernández Ballesteros, R. (1993). *Evaluación Psicológica.* Tomos 1 y 2. Madrid: Pirámide.
Forns, M.; Kirchner, T. & Torres, M. (1991). *Modelos de evaluación psicológica.* Barcelona: Barcanova.
Forns Santacana, M. (1993). *Evaluación psicológica infantil.* Barcelona: Barcanova.
García Cueto, E. (1993). *Introducción a la psicometría.* Madrid: Siglo XXI.





Garrett, H. E. (1937). On the interpretation of the standard error of measurement. *American Journal of Psychology. 49:* 679-680.
Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5:* 3 – 8.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6:* 427 – 439.
Hair, J. F. Jr.; Anderson, R. E.; Tatham, R. L. & Black, W. C. (1999). *Análisis multivariante.* Madrid: Prentice Hall.
Hammer, E. (1957). *Tests proyectivos gráficos.* Buenos Aires: Paidós.
Henderson, N. & Milstein, M. (2003). *Resiliencia en la escuela.* Buenos Aires: Paidós.
Hernández Sampieri, R.; Fernández Collado, C. & Baptista Lucia, P. (2000). *Metodología de la Investigación.* México, Mc Graw Hill.
Hogan, T. P. (2004). *Pruebas psicológicas. Una introducción práctica.* México: Manual Moderno.
International Test Commission (1996). *International Standards for Test Use.* Recuperado de http://www.intestcom.org/ITCguidelines/php, junio 4 de 2005.
International Test Commission (2006). *The ITC 5th International Conference on Psychological and Educational Test Adaptation across Language and Cultures. Building Bridges Among People.* Bruselas, 6 al 8 de julio: ITC.
Kaufman A. & Lichtenberger E. (1999). *Claves para la evaluación con el WAIS III.* Madrid: TEA
Koppitz, E. (1971). *El Test Guestáltico Visomotor para Niños.* Buenos Aires: Guadalupe.
Koppitz, E. (1995). *El Test de Bender.* Barcelona: Oikos-tau.
Lindzey, G. (1961). *Projective techniques and cross-cultural research.* New York: Invington.
Linley, P. A. & Joseph, S. (Eds.). (2004). *Positive psychology in practice.* New Jersey: Wiley.
Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Boston: Addison Wesley.
Machover, K. (1949). *Personality projection in the Drawing of the Human Figure.* Springfiled: Charles C. Thomas.
Maddux, J. E. (2002). Stopping the madness: Positive Psychology and the deconstruction of illness ideology and DSM. En C. R. Snyder & S. J. Lopez (Eds), *Handbook of positive psychology* (Cap. 2). New York: Oxford University Press.
Maganto, C.; Amador, J. A. & González, R. (2001). *Evaluación psicológica en la infancia y la adolescencia.* Madrid: TEA.
Martínez Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35:* 1012-1027.
Millon T. (1994). *Millon Clinical Multiaxial Inventory-III, (MCMI-III).* Minneapolis: National Computer Systems.
Millon, T. (1997). *El inventario de Estilos de Personalidad. MIPS.* Bs. Aires: Paidós.
Mosier, C. I. (1947). A critical examination of the concepto of face validity. *Educational and Psychological Measurement, 7:* 191-205.
Nahoum, C. (1961). *La entrevista psicológica.* Buenos Aires: Kapelusz.
O.M.S. (2003). Actualización de la Clasificación Internacional de Enfermedades, Décima Revisión (CIE-10). *Boletín Epidemiológico, 24(2),* junio.
Osgood, C. E.; Suci, G. C. & Tannenbaum, P. H. (1957). *The measurement of meaning.* Urbana: University of Illinois.
Pagano, R. (2006). *Estadística para la ciencias del comportamiento.* México: Thomson.
Pardo, A. & San Martín, R. (1998). *Análisis de datos en Psicología II.* Madrid: Pirámide.
Peterson, C. & Seligman, M.P. ( 2004). *Character strengths and virtues: A handbook and classification.* Washington DC: American Psychological Association.
Phillipson, H. (1983). Una breve introducción al Test de Relaciones Objetales. En R. Frank de Verthelyi (Ed.). *Actualizaciones en el Test de Phillipson* (cap.1), Buenos Aires: Paidós.

Pulido San Román, A. (1992): Estadísticas y técnicas de investigación social. Madrid: Pirámide.
Rappaport, D. (1978). *El modelo Psicoanalítico, la teoría del pensamiento y las técnicas proyectivas*. Buenos Aires: Paidós.
Real Academia Española (2001). *Diccionario de la Lengua Española*. Madrid: Espasa.
Rogers, C. (1966). *Psicoterapia centrada en el cliente*. Buenos Aires: Paidós.
Rohner, R. (1984). Towards a conception of culture for cross-cultural psychology. *Journal of Cross- Cultural Psychology, 15*, 111-138.
Rolla, E. (1981). *La entrevista en psiquiatría, psicoanálisis y psicodiagnóstico*. Buenos Aires: Galerna.
Rorschach, H. (1921/1942). *Psychodiagnostics: A diagnostic test based on perception*. Berna: Huber.
Russell, M.T. & Karol D.L. (2000). *16 PF-5*.Madrid: TEA.
Santisteban Requena, C. (1990). *Psicometría. Teoría y práctica en la construcción de tests*. Madrid: Norma.
Seligman, M.E.; Steen, T.; Park, N. & Peterson, C. (2005). Positive Psychology Progress. Empirical validation of interventions. *American Psychologist*, 60, (5), 410-421.
Silva, F. (1990). *Notas sobre el concepto de evaluación Psicológica. Comunicaciones*. Madrid: Colegio Oficial de Psicólogos.
Siquier de Ocampo, M.; García Arzeno, M. & Grassano, E. (1987). *La técnicas proyectivas y el proceso psicodiagnóstico*. Buenos Aires: Nueva Visión.
Snynder, C. R. & López, S. H. (2002). *Handbook of Positive Psychology*. New York: Oxford University Press.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677 – 680.
Sullivan, H. S. (1959). *La entrevista psiquiátrica*. Buenos Aires: Psiqué.
Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56* (3), 621-663.
Tellegen, A., & BenPorath, YS (1992). The new uniform T scores for the MMPI2: Rationale, derivation, and appraisal. *Psychological Assessment, 4*, 145-155.
Tavella, N. (1964): *Los test en la escuela*. Buenos Aires: Eudeba.
Wechsler, D. (1995). *La escala de inteligencia infantil WISC-III*. Buenos Aires: Paidós.
Wechsler, D. (1999). *WAIS – III. Escala de Inteligencia Wechsler para Adultos. Tercera Edición*. Madrid: TEA.
Wechsler, D. (2002). *WAIS – III. Test de Inteligencia para Adultos*. Buenos Aires: Paidós.
Wechsler, D. (2005). *WISC – IV. Escala de Inteligencia Wechsler para niños. Cuarta Edición*. Madrid: TEA.
Yela, M. (1987). *Apuntes de Psicología Matemática II*. Madrid: Facultad de Psicología, Universidad Complutense.
Zazzo, R.; Gilly, M. , Verba-Rao, M. (1970): *Nueva Escala Métrica de la Inteligencia*. Buenos Aires: Kapelusz).

# Índice

2-8-8-8